stichting
mathematisch
centrum

AFDELING MATHEMATISCHE STATISTIEK SN 7/77 NOVEMBER

J.G. BETHLEHEM, H. ELFFERS, R.D. GILL & J. RIJVORDT

# METHODEN, VOETANGELS EN KLEMMEN IN DE FACTORANALYSE

2e boerhaavestraat 49 amsterdam

*Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.*

*The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O).*

---

AMS(MOS) subject classification scheme (1970): 62H25

---

Methoden, voetangels en klemmen in de factoranalyse

door

J.G. Bethlehem, H. Elffers, R.D. Gill & J. Rijvordt

SAMENVATTING

Factoranalyse is een techniek voor de analyse van multivariaat waarnemingsmateriaal. Dit rapport behandelt een aantal vragen, die beantwoording behoeven, wil factoranalyse in een onderzoek van nut kunnen zijn.

TREFWOORDEN: *factoranalyse, principale componentenanalyse, datareductie*

INHOUD:

VOORWOORD

Het rapport "Methoden, voetangels en klemmen in de factoranalyse" is het product van een werkgroep van de afdeling Mathematische Statistiek van het Mathematisch Centrum in Amsterdam. De leden van deze groep hadden ieder voor zich aan hun ervaringen op het gebied van factoranalyse bij de statistische consultatie aan wetenschapsmensen uit tal van disciplines een gevoel van onvrede overgehouden. Zij besloten de onderhuids aanwezige gevoelens dat er bij veel toepassingen iets niet klopt, expliciet te maken.

Daarbij kwam al direct het vermoeden naar voren dat het merendeel der veronderstelde gebreken niet zozeer schuilt in de wiskundige inadequaatheid van de methode, als wel ligt op het moeilijke terrein van de aansluiting van het wiskundige model aan wat een onderzoeker met de methode hoopt te bereiken.

Van de aanvang af is het niet de bedoeling geweest om factoranalyse nietsontziend de grond in te boren, maar om te achterhalen wanneer deze methode zijn kracht kan ontplooien en wanneer niet. De schrijvers zijn daarbij op talloze twijfelachtigheden gestoten, d.w.z. momenten in de uitvoering van de analyse waarop de onderzoeker zeer alert moet zijn om niet zichzelf en zijn gehoor te misleiden met schijnbaar krachtige resultaten die in feite nietszeggend zijn.

Naar het de schrijvers voorkomt, is er sprake van een blinde vlek, een onopgevuld gebied, in de wijze van kijken naar factoranalyse, die niet omvat wordt door de wijze van kijken van wiskundigen, psychometrici of toepassers. Er zijn wiskundigen die op het standpunt staan dat alle moeilijkheden met een wiskundig kloppend model buiten hun verantwoordelijkheid vallen. Er zijn ook wiskundigen die vinden dat je alleen met een model mag werken als voor de volle 100 % de nodige veronderstellingen correct bevonden zijn. Er zijn toepassers die denken dat alles wat wiskundig geformuleerd is, boven kritiek verheven is, en anderen die menen dat alle kritiek van wiskundigen als academische scherpslijperij terzijde kan worden geschoven. De pur sang factoranalytici onder de psychometrici zijn veelal bezig met hogere orde verbeteringen van de factoranalysetechniek (zij ontwerpen bijvoorbeeld nog meer geavanceerde rotatietechnieken) en zijn van mening

dat factoranalyse als zodanig zijn nut wel bewezen heeft en dat de uitgangspunten geen kritische beschouwing meer behoeven.

Geen van deze vijf standpunten leidt tot een kritische bezinning op de aansluiting van het factoranalysemodel bij wat men er in de praktijk van verwacht, terwijl de schrijvers denken dat dat in iedere analyse opnieuw bijzonder noodzakelijk is.

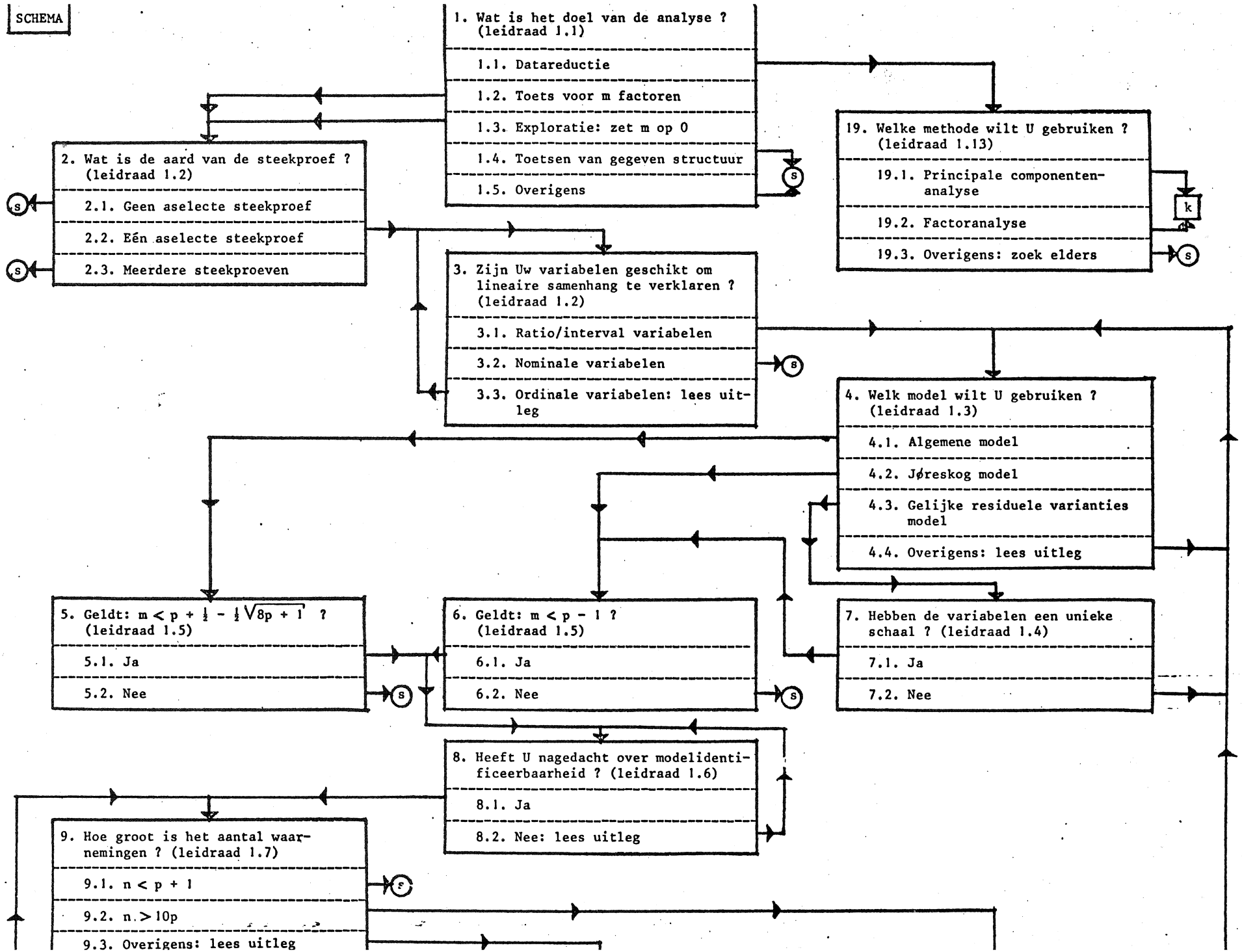
Alle bekende boeken over factoranalyse gaan mank aan hetzelfde gebrek: zij vermelden de nagestreefde doeleinden en geven weer wat er zoal wordt berekend, maar laten de lezer in de steek als het erop aankomt te beoordelen in hoeverre de nagestreefde doeleinden ook bereikt zijn.

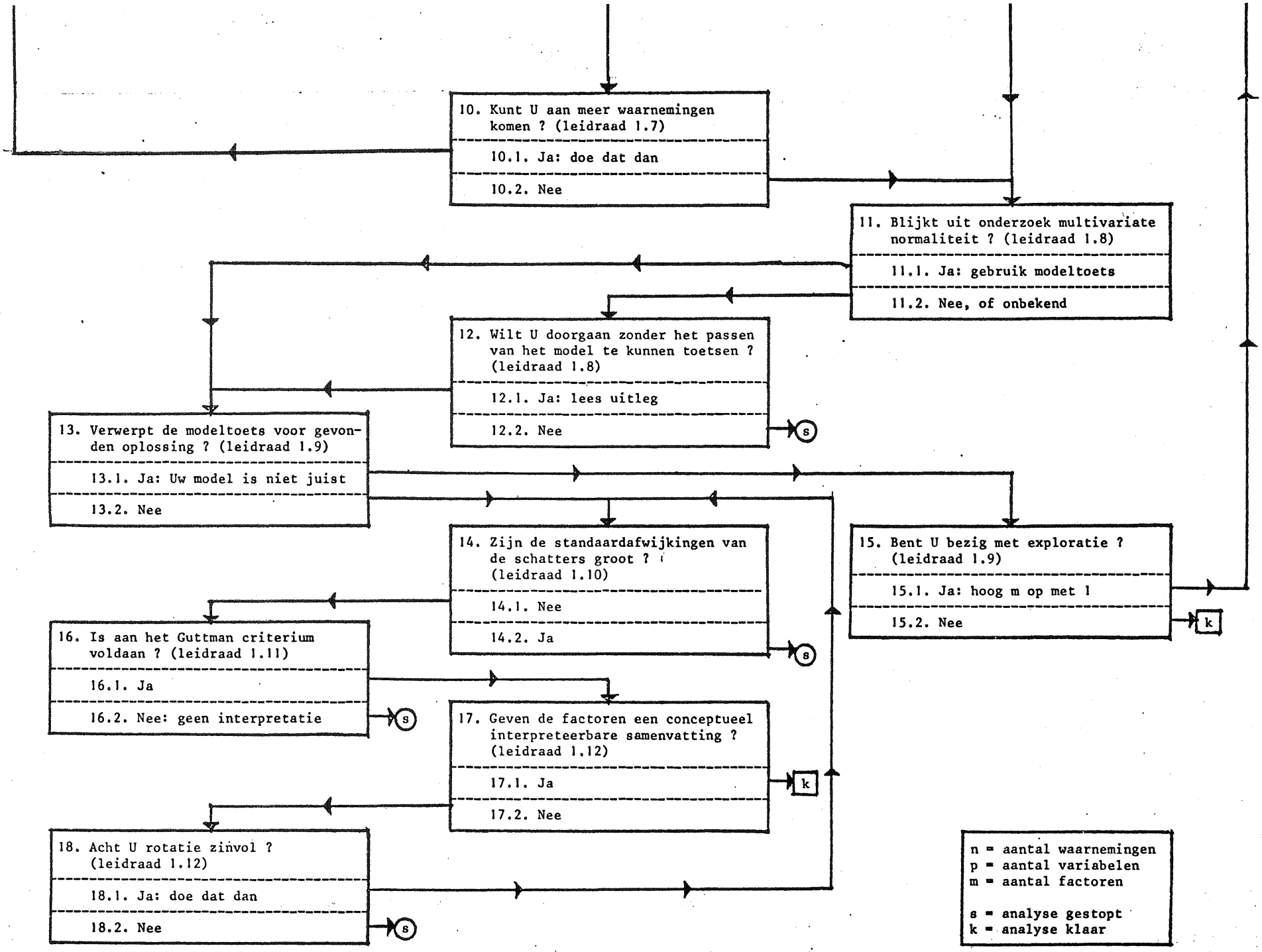
De schrijvers menen dat er te weinig wordt nagedacht over wat voor resultaat men nu eigenlijk heeft bereikt, wanneer aannemelijk is gemaakt dat een factorstructuur aan zeker materiaal ten grondslag kan worden gedacht en dat er verscheidene redenen zijn om een eenmaal gevonden factorladingenmatrix niet te kunnen interpreteren. De schrijvers geloven dat men bij de keuze voor, de uitvoering van en de resultateninterpretatie van factoranalyse tal van punten tegenkomt, waarbij men hoort stil te staan, om slechts na ampele overweging verder te gaan.

In dit rapport wordt getracht een systematische behandeling te geven van deze punten, die overigens meer algemeen methodologisch dan wiskundig-statistisch van aard zijn. Daarbij hebben de schrijvers hun best gedaan het geheel ook leesbaar te maken voor niet-statistici, in de hoop een bijdrage te leveren aan het dichten van de kloof tussen wiskundigen en gebruikers van de factoranalyse en aan een meer doordacht gebruik van deze techniek.

Tot slot kan nog opgemerkt worden dat de schrijvers, die hopen op opmerkingen en kritiek, dit rapport beschouwen als discussiestuk èn als handleiding voor de gebruikers van factoranalyse.

SCHEMA


1. Wat is het doel van de analyse ? (leidraad 1.1)
1.1. Datareductie
1.2. Toets voor m factoren
1.3. Exploratie: zet m op 0
1.4. Toetsen van gegeven structuur
1.5. Overigens
2. Wat is de aard van de steekproef ? (leidraad 1.2)
2.1. Geen aselecte steekproef
2.2. Eén aselecte steekproef
2.3. Meerdere steekproeven
19. Welke methode wilt U gebruiken ? (leidraad 1.13)
19.1. Principale componenten-analyse
19.2. Factoranalyse
19.3. Overigens: zoek elders
3. Zijn Uw variabelen geschikt om lineaire samenhang te verklaren ? (leidraad 1.2)
3.1. Ratio/interval variabelen
3.2. Nominale variabelen
3.3. Ordinale variabelen: lees uitleg
4. Welk model wilt U gebruiken ? (leidraad 1.3)
4.1. Algemene model
4.2. Jøreskog model
4.3. Gelijke residuele varianties model
4.4. Overigens: lees uitleg
5. Geldt: $m < p + \frac{1}{2} - \frac{1}{2}\sqrt{8p + 1}$ ? (leidraad 1.5)
5.1. Ja
5.2. Nee
6. Geldt: $m < p - 1$ ? (leidraad 1.5)
6.1. Ja
6.2. Nee
7. Hebben de variabelen een unieke schaal ? (leidraad 1.4)
7.1. Ja
7.2. Nee
8. Heeft U nagedacht over modelidentificeerbaarheid ? (leidraad 1.6)
8.1. Ja
8.2. Nee: lees uitleg
9. Hoe groot is het aantal waarnemingen ? (leidraad 1.7)
9.1. $n < p + 1$
9.2. $n > 10p$
9.3. Overigens: lees uitleg
s
k


10. Kunt U aan meer waarnemingen komen ? (leidraad 1.7)
10.1. Ja: doe dat dan
10.2. Nee
11. Blijkt uit onderzoek multivariate normaliteit ? (leidraad 1.8)
11.1. Ja: gebruik modeltoets
11.2. Nee, of onbekend
12. Wilt U doorgaan zonder het passen van het model te kunnen toetsen ? (leidraad 1.8)
12.1. Ja: lees uitleg
12.2. Nee
s
13. Verwerpt de modeltoets voor gevonden oplossing ? (leidraad 1.9)
13.1. Ja: Uw model is niet juist
13.2. Nee
14. Zijn de standaardafwijkingen van de schatters groot ? (leidraad 1.10)
14.1. Nee
14.2. Ja
s
15. Bent U bezig met exploratie ? (leidraad 1.9)
15.1. Ja: hoog m op met 1
15.2. Nee
k
16. Is aan het Guttman criterium voldaan ? (leidraad 1.11)
16.1. Ja
16.2. Nee: geen interpretatie
s
17. Geven de factoren een conceptueel interpreteerbare samenvatting ? (leidraad 1.12)
17.1. Ja
17.2. Nee
k
18. Acht U rotatie zinvol ? (leidraad 1.12)
18.1. Ja: doe dat dan
18.2. Nee
s
n = aantal waarnemingen
p = aantal variabelen
m = aantal factoren
s = analyse gestopt
k = analyse klaar

A factor problem starts with the hope or conviction that a certain domain is not so chaotic as it looks.

-L.L. Thurstone, Multiple Factor Analysis (1947)-

# INLEIDING

Factoranalyse is een veel gebruikte techniek bij het verwerken van multivariaat waarnemingsmateriaal. Het beantwoorden van de vraag of, en zo ja, hoe, factoranalyse in een bepaald onderzoek van nut kan zijn, is niet eenvoudig. Misschien is dat de reden dat die vraag soms in het geheel niet wordt gesteld. Ook het beoordelen van de bruikbaarheid van de resultaten van een factoranalyse vereist zorgvuldige overweging.

Het is de mening van de schrijvers van dit rapport dat onvoldoende aandacht voor deze vragen maar al te vaak oorzaak is van resultaten die de toets der kritiek niet kunnen doorstaan. Het rapport wil aangeven waarom bepaalde vragen noodzakelijk antwoord, of in ieder geval overweging, behoeven, wil factoranalyse een zinvolle techniek zijn. Het rapport is geen gedetailleerde handleiding hoe men factoranalyse met behulp van een computerprogramma uitvoert, hoewel rekenkundige kwesties zeker aan de orde komen, doch het behandelt vooral wat men voor en na het eigenlijke rekenwerk dient te doen.

Het rapport bestaat uit drie delen, die naar bij de lezer verondersteld mathematisch niveau uiteenlopen:

a. Het eerste deel is de LEIDRAAD. In de leidraad wordt de gebruiker van factoranalyse een groot aantal samenhangende vragen gesteld en, afhankelijk van de antwoorden, adviezen gegeven. De leidraad is samengevat in een SCHEMA. De leidraad vraagt van de lezer een elementaire kennis van de statistiek, i.h.b. van correlatierekening.
b. Het tweede deel, de UITLEG, belicht een aantal situaties nader, waarover de leidraad vragen stelt en verduidelijkt het hoe en waarom van deze vragen. Naast de kennis, benodigd voor het lezen van de leidraad, is een zekere vertrouwdheid met matrixrekening nuttig voor sommige delen van de uitleg.
c. Het derde deel bevat de DETAILS: een aantal bewijzen en omvangrijkere voorbeelden. Het wiskundig niveau is uiteenlopend, maar eist soms een gedegen kennis van de multivariate mathematische statistiek.

Wanneer men niet geïnteresseerd is in de achtergronden van de adviezen, kan men de leidraad ook afzonderlijk lezen, mits men over een zekere bekendheid met factoranalytische technieken beschikt.

# 1. LEIDRAAD

## 1.0. INLEIDING

Deze leidraad tracht de factoranalyticus te dwingen tot een antwoord op een aantal vragen omtrent zijn bedoelingen en zijn (geplande) waarnemingsmateriaal. Hij geeft adviezen of stelt vragen naar aanleiding van de antwoorden.

Er zijn drie typen vragen: naar aard en bedoeling van het onderzoek, over keuze en specificatie van een factormodel en over het beoordelen van de bruikbaarheid van de resultaten.

Om de onderzoeker er van te doordringen dat het niet gaat om academische kwesties over factoranalyse, maar om het nut van factoranalyse voor zijn onderzoek, is de leidraad in de "U-stijl" geschreven. Een korte samenvatting van de leidraad vindt U in het SCHEMA.

Uitgangspunt van deze leidraad is dat U

a. aan een aantal experimentele grootheden (personen, objecten, situaties etc.) een aantal metingen zult verrichten (of eventueel al hebt verricht), en wel aan elke eenheid dezelfde metingen;
b. aan factoranalyse denkt als een mogelijke nuttige analysemethode voor Uw onderzoek.

Als bovenstaande niet het geval is, zal het niet zinvol zijn in dit rapport te rade te gaan.

In dit rapport wordt rijkelijk gebruik gemaakt van bekende begrippen uit de factoranalyse, zoals *factoren*, *factorladingen*, *uniciteiten*, *communaliteiten*, *factorscores* e.d. In 2.1. worden al deze begrippen geïntroduceerd. U kunt deze sectie dan ook beter eerst lezen als U niet vertrouwd bent met deze begrippen of om te zien hoe wij deze begrippen precies gebruiken. Wij raden U aan zonodig af en toe 2.1. na te slaan als U niet begrijpt waarover vragen en antwoorden gaan. In 2.2. vindt U een symbolenverklaring.

## 1.1. DOEL VAN DE ANALYSE

Waarom dacht U aan factoranalyse? Wat wilt U ermee voor Uw onderzoek? Op deze vragen zijn, in deze context, de volgende antwoorden van toepassing:

a. U beoogt *datareductie* tot stand te brengen.
b. U hoopt een *factorstructuur* te *ontdekken*.
c. U wilt een gegeven *factorstructuur* toetsen.
d. U heeft *iets anders* op het oog.

*Iets anders*

Om met het laatste te beginnen: De auteurs van dit rapport kunnen zich niet voorstellen dat factoranalyse voor iets anders dan datareductie of het vinden of toetsen van een factorstructuur nuttig kan zijn. U kunt dus niet in dit rapport terecht. (Dit houdt niet in dat andere multivariate technieken zoals bijvoorbeeld één- of tweesteekproeventoetsen, multivariate variantieanalyse, canonieke correlatieanalyse, discriminantanalyse of clusteranalyse niet zinvol kunnen zijn. Zij vallen echter buiten het bestek van dit rapport.)

Om te kunnen kiezen uit de overige drie antwoorden moet U natuurlijk weten wat een factorstructuur is. Zonodig kunt U 2.1. raadplegen. Wat we onder datareductie verstaan, wordt hieronder uiteengezet.

*Datareductie*

Wanneer U, alvorens een of andere bewerking op het materiaal uit te voeren, dit materiaal overzichtelijker wilt maken door het aantal metingen per experimentele eenheid dat U in ogenschouw neemt, te verminderen, dan spreken wij van datareductie. Natuurlijk is het beter niet eerst meer metingen te produceren dan men eigenlijk kan of wil verwerken.
Dit kan vaak door adequate planning vermeden worden. Toch kan het gebeuren, en in feite komt het veel voor, dat behoefte aan datareductie ontstaat, bijvoorbeeld omdat de materiaalverzameling voor een meeromvattend doel heeft plaatsgevonden dan voor de analyse waar U nu mee bezig bent.

Kenmerken van datareductie zijn:

a. Het gebeurt niet om zich zelfs wille, maar om verdere analyse te vergemakkelijken.
b. Er is geen pretentie dat op ander materiaal dezelfde manier van reductie behoort te worden toegepast, of succesvol zal zijn.
c. Elke methode die leidt tot reductie van het aantal kenmerken per eenheid is in principe geslaagd.

Vaak hoopt men "niet te veel informatie te verliezen". Afhankelijk van wat men daarmee bedoelt, komen bepaalde methodes van datareductie meer of minder in aanmerking. (Zie hiervoor 1.13.) Een methode van datareductie is bijvoorbeeld het kiezen van een beperkt aantal metingen.

Wat heeft factoranalyse eigenlijk met datareductie te maken? Historisch werd het onderscheid tussen datareductie en het ontdekken van een factorstructuur in de zin van de volgende alinea niet gemaakt. Daar factoranalyse de aangewezen weg is om een factorstructuur te ontdekken, leek het dus ook de aangewezen weg voor datareductie. Verder wordt een veelgebruikte analysemethode voor datareductie, de *principale componentenanalyse*, ten onrechte voor het ontdekken van een factorstructuur gebruikt. Wij zullen principale componentenanalyse streng onderscheiden van factoranalyse in eigenlijke zin. De verschillen worden uiteengezet in 2.8.

Bent U uit op datareductie, dan zijn er vele mogelijkheden. Misschien hebt U nog iets aan de overwegingen in 1.13. over principale componentenanalyse als datareductiemethode. Het is niet onmogelijk factoranalyse in eigenlijke zin voor datareductie te gebruiken, maar het is in de praktijk zelden een gelukkige keus. Verspreid in dit rapport staan opmerkingen over factoranalyse als datareductiemethode, hoewel het rapport zich in het algemeen richt op factoranalyse als middel voor structuuronderzoek. Overigens hebt U voor datareductie niet veel aan dit rapport.

*Ontdekken van een factorstructuur*

Bij het trachten te vinden van een factorstructuur gaat het om het volgende: U poogt een theorie over een of ander gebied van waarneembare verschijnselen te ontwikkelen en U bent er in het bijzonder op gespitst een verklaring te geven voor de samenhang van bepaalde metingen aan expe-

rimentele eenheden. U acht het redelijk te postuleren dat samenhang tussen elk tweetal waargenomen variabelen voortkomt uit hun gedeeltelijk bepaald zijn door een kleiner aantal andere, niet waargenomen variabelen. (Zie 2.7. voor een aantal overwegingen hierbij.) Het ontdekken van een factorstructuur is nu er achter komen in welke mate en hoe deze bepaling door nieuwe variabelen bestaat. Deze nieuwe variabelen heten factoren. We onderscheiden twee varianten in het ontdekken van een factorstructuur:

a. *de* m-*factorvariant*
   U wilt *aantonen* dat de samenhang tussen Uw variabelen wordt veroorzaakt door een bepaald, vast, door U van te voren gespecificeerd aantal, m, factoren, en U wilt *nagaan* op welke wijze die factoren de variabelen bepalen.
b. *de exploratieve variant*
   U wilt nagaan of het mogelijk is met een niet van te voren gespecificeerd aantal factoren de samenhang tussen Uw variabelen te verklaren. U wilt daarbij zowel het aantal factoren, waarmee dit (eventueel) mogelijk is, vaststellen, alsook de wijze waarop zij de variabelen bepalen, nagaan.

Kenmerken van het structuurontdekken zijn:

a. Het gebeurt, binnen het kader van een theorie, om zich zelfs wille, althans in eerste instantie;
b. Er bestaat de pretentie dat de gevonden resultaten, zo die er zijn, ook voor nieuwe metingen op dit gebied van toepassing zijn;
c. Niet alle wiskundig voldoende oplossingen worden als succesvol beschouwd; zij moeten ook in overeenstemming zijn met de theorie in kwestie en verklarende kracht hebben.

Wanneer U uit bent op het ontdekken van een factorstructuur, stel dan vast welke van beide varianten U op het oog hebt en lees dan verder in 1.2.

*Toetsen van een factorstructuur*

Als U uit eerder onderzoek, of hoe dan ook, een veronderstelling hebt over het bestaan van een welomschreven factorstructuur, dan kunt U die ver-

onderstelling willen toetsen. De gang van zaken bij het toetsen op aanwezigheid van een bepaalde factorstructuur is in dit rapport niet beschreven. Voor een uiteenzetting over dit onderwerp verwijzen we naar LAWLEY & MAXWELL (1971). Niettemin zijn vele overwegingen die in dit rapport de revue passeren ook in dit kader steekhoudend.

## 1.2. AARD EN HERKOMST VAN DE VARIABELEN

Factoranalyse en principale componentenanalyse zijn beide technieken om een covariantie- of correlatiematrix te onderzoeken. Hoewel het, gegeven een n-tal experimentele eenheden aan elk waarvan p metingen zijn verricht, altijd formeel mogelijk is een covariantiematrix uit te rekenen en deze aan een of andere analysetechniek te onderwerpen, trekt U slechts profijt van de hele statistische machinerie als deze covariantiematrix zijn gewone betekenis heeft, d.w.z. als hij beschouwd kan worden als een schatting voor de populatie covariantiematrix. Daartoe is het noodzakelijk dat U de n experimentele eenheden kunt beschouwen als n onafhankelijke (aselecte) trekkingen uit één en dezelfde populatie, of anders uitgedrukt: dat het n-tal p-voudige metingen n onafhankelijke realisaties van één p-dimensionale stochastische vector zijn.

Beschouwen wij allereerst de frase "trekkingen uit een en dezelfde populatie". In welk van de onderstaande drie gevallen bevindt U zich?

a. U wilt of U kunt Uw data misschien helemaal niet zien als waarnemingen aan een stochastische vector. U hebt dan weinig aan dit rapport. U bent mogelijk bezig met beschrijvende statistiek, of U bevindt zich toch in de "alles mag"-situatie van datareductie.
b. U bent meer geneigd Uw waarnemingen te beschouwen als realisaties van twee of meer verschillende stochastische vectoren, bijvoorbeeld omdat U waarnemingen verricht bij een experimentele en een controle groep. In dat geval is het vrijwel onmogelijk zinnig gebruik te maken van factoranalyse. Lees 2.6. en beëindig het lezen van dit rapport, tenzij U zich in één der daar aangestipte bijzondere gevallen bevindt.
c. U meent Uw data wel te kunnen beschouwen als n trekkingen uit dezelfde populatie. Dan kunt U wel een echte covariantiematrix schatten, als U tenminste ook nog die trekkingen als aselect kunt beschouwen. Is dit

niet het geval, dan kunt U hier niet van steekproef naar populatie generaliseren. Een voorbeeld: stel dat er één belangrijke factor achter de p variabelen in de populatie zit, maar Uw steekproefsysteem neemt alleen experimentele eenheden, die een bepaalde waarde van die factor hebben, in de steekproef op. Dan zult U die factor niet vinden. Bovendien is de interpretatie van een dan eventueel gevonden factor erg moeilijk: U kunt niet generaliseren naar de populatie.

Dus alleen als U over een aselecte steekproef uit één en dezelfde populatie beschikt, kunt U op grond van Uw steekproef statistische uitspraken doen over een factorstructuur in die populatie - zie voor een nadere toelichting hierover 2.5. Bedenkt U dan nog dat U bij Uw analyse uitgaat van de covarianties tussen de verschillende variabelen en dat in het algemeen covarianties alleen geschikte samenhangsmaten zijn voor variabelen, die, gemeten op een intervalschaal, een lineaire samenhang vertonen.

Wanneer de variabelen in feite een minder eenvoudig type samenhang vertonen, dan vindt U dat meestal niet. (Zie 2.7. voor een andere uiteenzetting.) Realiseert U zich tenslotte dat het gebruik van een covariantiematrix in plaats van de originele waarnemingen, behoudens bijzondere gevallen, (bijv. als Uw waarnemingen een multivariate normale verdeling hebben) tot een groot informatieverlies kan leiden. (Zie eveneens 2.7.) Als ook dit voor U geen probleem vormt, lees dan door in 1.3.

## 1.3. MODELKEUZE

Deze paragraaf vraagt U welk model U op het oog heeft. Daarmee samenhangend wordt U ook een keuze van de oplossingsmethode aangeraden. Bij sommige factormodellen zijn namelijk verschillende oplossingsmethoden voorgesteld met zeer uiteenlopende kwaliteiten. U kiest voor:

a. *het algemene of traditionele model*
   Dit gebruikt de methode der meest aannemelijke schattingen (onder normaliteit), gevonden met de methode van Jøreskog voor dit model.
   Lees door in 1.5.
b. *het Jøreskogmodel*
   Dit is het algemene model, uitgebreid met een extra aanname over de vorm

van de uniciteiten van de variabelen. Er is een directe oplossingsmethode. Lees door in 1.5.

c. *het gelijke residuele variantiesmodel*
Ook dit is een variatie op het algemene model, verkregen door een andere aanname over de vorm van de uniciteiten. De oplossingsmethode staat beschreven in 3.1. Lees door in 1.4.

d. U weet het niet, of U had een andere methode op het oog. Leest U eerst de uitleg in 2.3.

## 1.4. UNIEKE SCHAAL

Daar U het gelijke residuele variantiesmodel van toepassing acht, is de volgende vraag van wezenlijk belang:

Hebben de variabelen een unieke schaal? D.w.z. maakt het in Uw ogen iets uit als U de schaal van een of meer variabelen zoudt veranderen? Als Uw antwoord hierop "neen" luidt, dus als U geen positieve redenen hebt om nu juist alle variabelen in een bepaalde schaal uit te drukken, dan heeft het geen zin om door te gaan met dit model. Een schaalverandering kan radicaal andere resultaten geven, terwijl U voor schaalveranderingen onverschillig zegt te zijn. Als Uw materiaal aan de modelveronderstelling voldoet, dan kan dat door schaalveranderingen ongedaan gemaakt worden en omgekeerd. (Een nadere uiteenzetting over het effect van schaalveranderingen vindt U in 2.9.)

Een en ander geldt niet voor collectieve schaalverandering: als U alleen onverschillig bent tegenover een gelijke schaalverandering voor alle variabelen tegelijk, dan is dàt niet de doodssteek voor het gelijke residuele variantiesmodel.

Vaak wordt getracht dit bezwaar te omzeilen door alle variabelen zo te schalen dat ze variantie 1 hebben, d.w.z. dat de analyse wordt toegepast op de correlatiematrix i.p.v. op de covariantiematrix. In dit geval houdt toepasbaarheid van het gelijke residuele variantiesmodel de volgende veronderstelling in: elke variabele wordt voor een even groot gedeelte bepaald door de factoren. Wilt U dit volhouden dan moet U daar ook een positieve reden voor hebben.

Beantwoordt U de vraag naar uniekheid van de schaal bevestigend, gaat U dan door bij 1.5. In het andere geval raden wij U aan U te bezinnen op een ander model.

## 1.5. AANTAL FACTOREN

Als U de samenhang van een p-tal variabelen hoopt te verklaren met behulp van een m-tal andere variabelen, factoren, dan is het zaak aandacht te besteden aan de grootte van m ten opzichte van p. Voor te grote m namelijk is de uitspraak "m factoren zijn voldoende om de samenhang van deze p variabelen te verklaren" geen interessante uitspraak meer: hij is voor elk p-tal variabelen waar en zegt niets over deze variabelen in het bijzonder. De uitspraak vormt geen restrictie op de samenhang van de variabelen. (Zie voor een nadere toelichting 2.4.)

Voor de verschillende modellen die hier behandeld worden, ligt de grens van het aantal factoren dat nog tot een interessante uitspraak leidt, verschillend. Voor het algemene model is er zelfs geen grens, maar slechts een vuistregel ter beschikking:

Geldt, terwijl U met de m-factor variant (1.1.) bezig bent, voor het aantal factoren m:

a. in het algemene model: $m < p + \frac{1}{2} - \frac{1}{2}\sqrt{8p + 1}$?
b. in het Jøreskog-model: $m < p - 1$?
c. in het gelijke residuele varianties-model: $m < p - 1$?

Zo ja, vervolg dan deze leidraad in sectie 1.6. Zo nee, dan bent U beland in een situatie waar een factorstructuur niet interessant is als verklaring van samenhang. Hoogstens kunt U hem zien als een alternatieve beschrijving van samenhang, die evenwel niets nieuws toevoegt. Wij raden U aan te stoppen met Uw factoranalyse.

Hebt U het aantal factoren nog niet gespecificeerd, d.w.z. bent U bezig met de exploratieve variant (1.1., blz. 5), dan moet U door opeenvolgend proberen met verschillende waarden van m zien te bepalen welke de best passende waarde is, zo er tenminste één past. Deze procedure is niet zonder bezwaren, die nader beschreven staan in 2.10. U begint met voor m de volgens

U laagste waarde te kiezen. Mocht in de loop van de analyse, met name bij de vragen uit 1.9. blijken, dat dit niet tot een geschikte oplossing leidt, dan herhaalt U Uw exploratie met een waarde van m die één hoger is. Daarbij moet U wel telkenmale de vragen die in 1.9. omtrent de grootte van m worden gesteld, bevestigend kunnen beantwoorden. Is dit niet het geval, of vindt U de zo verkregen waarde van m te hoog om interessant te zijn, dan eindigt Uw exploratie hier met de conclusie dat het niet mogelijk is de samenhang tussen Uw variabelen met een voldoende klein aantal factoren te verantwoorden.

## 1.6. IDENTIFICEERBAARHEID

De vraag naar identificeerbaarheid van het factormodel is de vraag of er precies één oplossing is, zo er tenminste een oplossing bestaat. Wanneer er meerdere oplossingen bestaan, komt U in moeilijkheden met de interpretaties, de algoritmen die U een oplossing geven, krijgen soms zeer onaangename eigenschappen en U zult niet-statistische argumenten moeten vinden om een oplossing uit te kiezen. Er zijn altijd twee vormen van niet-identificeerbaarheid:

a. Met elke matrix van factorladingen die voldoet, voldoet ook elke orthogonale rotatie ervan. Aan dit probleem valt gedeeltelijk tegemoet te komen, zie 2.13.
b. Bij een gegeven oplossing bestaan altijd nog vele andere variabelen die de rol van een bepaalde factor kunnen vervullen. Dit probleem wordt in 1.11. apart behandeld.

Daarnaast kan het voorkomen dat verschillende matrices van factorladingen voldoen die niet door rotatie uit elkaar te verkrijgen zijn. De vraag of dit inderdaad zo is, is in het algemeen zeer moeilijk te beantwoorden voor het algemene model. Zeker is, dat het model niet altijd identificeerbaar is en dat in niet-identificeerbare gevallen vervelende verschillen kunnen optreden. Een nadere uiteenzetting vindt U in 2.4. Het Jøreskog-model en het gelijke residuele variantiesmodel zijn onder de conditie van 1.5. altijd identificeerbaar in deze derde zin.
Als U niet al te zeer ontmoedigd bent, vervolg dan de leidraad bij 1.7.

## 1.7. AANTAL WAARNEMINGEN

Het aantal experimentele eenheden n, waaraan telkens p waarnemingen zijn verricht, is van groot belang voor de nauwkeurigheid van de te verkrijgen oplossingen. In welke van de onderstaande gevallen bevindt U zich?

a. n *kleiner dan of gelijk aan* p
   In dit geval moet U de analyse stopzetten, daar U met dit aantal waarnemingen de covariantie- of correlatiematrix niet goed genoeg kunt schatten. Bovendien is meestal bij de oplossingsmethode de inverse van deze matrix nodig en deze bestaat niet, aangezien de covariantie- of correlatiematrix singulier is. Alleen als U aan meer waarnemingen kunt komen, is er nog hoop.
b. n *ongeveer* 10 *maal* p *of meer*
   Dit aantal is in het algemeen voldoende voor nauwkeurige oplossingen. Leest U door in 1.8.
c. n *ligt tussen* p *en* 10 *maal* p *in*
   Dit is een twijfelgeval. Het relatief geringe aantal waarnemingen kan leiden tot onnauwkeurige schattingen (zie 1.10.) en tot een onbetrouwbare modeltoets (zie 1.9.). Als het enigszins mogelijk is, probeert U dan aan meer waarnemingen te komen. Dit is des te meer nodig naarmate n dichter bij p ligt. Lukt het U niet aan meer waarnemingen te komen, gaat U dan toch maar verder bij 1.8., maar U moet niet vreemd opkijken als de verdere analyse teleurstellend verloopt.

## 1.8. VERDELING VAN DE WAARNEMINGEN

Hebben de p waarnemingen aan één experimentele eenheid, althans bij benadering, een p-variate normale verdeling? Als Uw antwoord hierop bevestigend is, bevindt U zich in een zeer aangename situatie: U kunt niet alleen de parameters in het model schatten, maar bovendien kunt U iets aan de weet komen omtrent de nauwkeurigheid van de schattingen en U kunt een toets uitvoeren of het model wel overeenkomt met Uw waarnemingsmateriaal. U kunt zelfs deze toets gebruiken om het aantal factoren te schatten als dat van te voren onbekend was. (Zie 1.5. en 2.10.) Wel zijn de procedures om de nauwkeurigheid van de schattingen en de passendheid van het model te

onderzoeken alleen geldig voor grote aantallen waarnemingen. LAWLEY & MAXWELL (1971) adviseren vertrouwen te hebben in de uitslag van een toets wanneer $n > p + 50$ . Ga door bij 1.9.

Gelooft U niet dat Uw waarnemingen uit een multivariate normale verdeling komen (ook niet bij benadering), dan heeft U weinig mogelijkheden om de resultaten van analyse naar waarde te schatten. Merk op dat dit de exploratieve variant (zie 1.1.) onmogelijk maakt. Wel zijn de schatters van de parameters in Uw model in het algemeen dan nog asymptotisch raak, maar het is niet mogelijk meer te weten te komen over de nauwkeurigheid van die schattingen, noch of het model wel past. Vindt U dat, ons inziens terecht, een te groot bezwaar, dan eindigt hier Uw analyse, althans van de gegevens in de huidige vorm. Soms kan een transformatie (logaritmisch bijv.) uitkomst brengen. Wilt U doorgaan, leest U dan eerst 2.11. alvorens verder te lezen bij 1.11.

## 1.9. UITSLAG VAN DE MODELTOETS

De modeltoets, die slechts kan worden uitgevoerd wanneer de waarnemingen uit een multivariate normale verdeling komen, toetst de hypothese: "m factoren zijn voldoende om de samenhang tussen de p variabelen te verklaren" tegen het alternatief dat er meer factoren nodig zijn.

Verwerpt de modeltoets de gevonden oplossing, dan concludeert U dat m factoren niet voldoende zijn. Als U redenen hebt om aan te nemen dat Uw hypothese toch juist is, dan moet U nog eens naar de verdeling van Uw waarnemingsmateriaal kijken, want ook niet-normaliteit kan de oorzaak van verwerpen van de nulhypothese zijn. Anders zijn Uw verwachtingen niet uitgekomen en stopt U hier met Uw analyse, tenzij U met de exploratieve variant bezig bent (1.1.), in welk geval U in het schema teruggaat naar 15 om de waarde van m met 1 op te hogen.

Verwerpt de toets niét, dan is het niet onredelijk te veronderstellen dat een m-factorstructuur aanwezig is. In 2.10. wordt echter uiteengezet dat de modeltoets in de exploratieve variant niet zo'n schitterende manier is om het juiste aantal factoren te bepalen. Wij raden U daarom ten sterkste aan om de zo gevonden structuur nogmaals op *nieuw* materiaal te

toetsen door daarop een analyse in de m-factor variant uit te voeren met als waarde van m het aantal in de exploratieve fase gevonden factoren.

Alle verdere overwegingen in de volgende paragrafen zijn slechts geldig wanneer U op deze wijze het juiste aantal factoren hebt bepaald. Tevens moet dan worden aangenomen dat het factormodel identificeerbaar is (1.6.). Gaat U verder in 1.10.

## 1.10. NAUWKEURIGHEID VAN DE SCHATTERS

Wanneer U tot deze paragraaf bent doorgedrongen, weet U dat Uw veronderstelling dat er m factoren bestaan die de lineaire afhankelijkheid tussen de variabelen verklaren, niet onredelijk is. Alle verdergaande interpretatie berust op factorladingen. Deze kent U niet, maar U heeft ze geschat. Het heeft alleen zin een verfijnde interpretatie op deze schattingen te baseren als U er op kunt vertrouwen dat deze schattingen enigszins nauwkeurig zijn.

Om een indruk hiervan te krijgen rekent U de geschatte standaardafwijkingen van de schatters van de factorladingen uit. Dit kan alleen als de waarnemingen uit een multivariate normale verdeling komen. Zijn deze standaardafwijkingen klein, dan lijkt het er op, dat U voldoende zekerheid hebt over de grootte van de factorladingen om Uw interpretatie er op te kunnen baseren.

Wanneer noemt U de geschatte standaardafwijking klein? Dat hangt af van de mate waarin U de grootte van de ladingen bij Uw interpretatie wilt laten meespelen. Als regel zou genomen kunnen worden: de standaardafwijkingen zijn klein genoeg, wanneer ze gemiddeld niet groter zijn dan de helft van het kleinste absolute verschil tussen de factorladingen dat U nog relevant acht bij interpretatie.

Voorbeeld 1: Bij interpretatie van factoren wilt U een lading van 0,6 anders behandelen dan een van 0,4. Het verschil is 0,2. De standaardafwijkingen van de schatters mogen dan gemiddeld niet groter zijn dan 0,1.

Voorbeeld 2: Bij interpretatie wilt U nog aandacht besteden aan de eerste decimaal van de factorladingen. Het minimaal te interpreteren verschil is dan dus 0,1. De standaardafwijkingen moeten dan gemiddeld kleiner zijn dan 0,05.

Voor vele gevallen is de eis dat de geschatte standaardafwijkingen

gemiddeld niet groter zijn dan 0,1 wel voldoende.

Voldoen de standaardafwijkingen aan de door U gestelde eis, ga dan door bij 1.11. Zijn deze standaardafwijkingen niet klein dan is de onzekerheid over de factorladingen te groot om te kunnen interpreteren. U moet Uw analyse overdoen met meer waarnemingen als U in betere resultaten geïnteresseerd bent. (Voor de berekening van standaardafwijkingen zie 3.2.)

## 1.11. GEDETERMINEERDHEID VAN DE FACTOREN

U weet nu redelijk nauwkeurig hoe groot de factorladingen zijn. Voordat U de factoren op grond van deze ladingen gaat benoemen, is het goed de zogenaamde *Guttman criteriumwaarde* (GCW) voor elke factor te berekenen. Het is namelijk niet zo, dat, gegeven de factorladingen, er slechts één nieuwe variabele (factor) bestaat die juist de factorladingen als correlatiecoëfficiënten met de oude variabelen heeft. In het algemeen zijn er zeer vele variabelen, zogenaamde kandidaat-factoren, die hieraan voldoen, terwijl deze variabelen onderling sterk kunnen verschillen. Wanneer er nu twee van deze kandidaat-factoren bestaan, die een lage correlatiecoëfficiënt hebben, dan is het duidelijk dat de interpretatie van de betekenis van deze factor op losse schroeven staat: bij elke interpretatie kan een andere, die vrijwel niets met de eerste te maken heeft, worden gegeven.

De GCW is de laagst bereikbare correlatiecoëfficiënt tussen twee kandidaten voor een factor. (Zie 2.12.) Ligt het criterium dicht bij 1, hetgeen bijvoorbeeld voorkomt als er zeer hoge ladingen bij die factor aanwezig zijn, dan lijken al de kandidaten voor die factor sprekend op elkaar. U kunt deze factor gaan interpreteren, leest U door in 1.12. Ligt de GCW in de buurt van 0 (of is zij zelfs negatief) dan kunnen kandidaat-factoren zelfs bijna (of helemaal) ongecorreleerd zijn. U kunt deze factor dan niet veilig benoemen. Merk op, dat ook de berekening van de factorscores zinloos wordt, wanneer de GCW laag is.

Wanneer U factoranalyse gebruikt t.b.v. datareductie, wilt U voor elke factor behalve een hoge GCW bovendien dat elke variabele een hoge communaliteit heeft. U wilt immers de factorscores gebruiken voor verdere analyse. Bij schatting van de factorscores spelen variabelen met een lage

communaliteit nauwelijks een rol; dat een variabele weinig samenhang vertoont met de andere, wil echter nog niet zeggen dat die variabele voor Uw doel onbelangrijk is.

Gebruikt U principale componentenanalyse, als datareductietechniek, dan zijn er criteria, analoog aan het communaliteitencriterium bij factoranalyse. U wilt, om te vermijden dat mogelijk belangrijke variabelen geheel weggereduceerd worden, dat alle variabelen voor een belangrijk gedeelte door de componenten worden bepaald. Daartoe zal het "percentage door de componenten verklaarde variantie" zeer groot moeten zijn. Hier is alles overigens veel subjectiever: Uw keuze van m componenten wordt al dan niet gerechtvaardigd door succes of mislukking van wat U er verder mee doet.

## 1.12. INTERPRETATIE VAN FACTOREN

U weet nu, dat het wiskundig niet onmogelijk is dat al Uw p variabelen bepaalde lineaire combinaties zijn van m nieuwe variabelen (de factoren) en de p specifieke gedeelten, en dat volgens het Guttmancriterium deze factoren vrij nauwkeurig gedetermineerd zijn.

De vraag is nu: zijn deze factoren ook te benoemen binnen de theorie waarmee U werkt? Kunt U deze factoren als meer zien dan een louter wiskundige formulering van de samenhang in Uw materiaal? Daartoe tracht U de factoren te interpreteren, d.w.z. ze een naam en betekenis te geven, op grond van hun correlatiecoëfficiënten met de originele variabelen. (Deze correlatiecoëfficiënten zijn juist de factorladingen, indien U bent uitgegaan van gestandaardiseerde variabelen.)

Vervolgens vraagt U zich af: is het bestaan van variabelen met deze namen en betekenissen rijmbaar met de theorie, en is het te verdedigen dat de geobserveerde variabelen van deze factoren afhangen? Het beantwoorden van deze vragen is tot op zekere hoogte subjectief en eist een hoge mate van integriteit van de onderzoeker. Het is in ieder geval aan te bevelen via andere wegen naar bevestiging van Uw resultaten te zoeken. Kunt U bovenstaande vragen bevestigend beantwoorden, dan hebt U de analyse daarmee tot een goed einde gebracht, in het andere geval bent U op het allerlaatst nog gestrand.

Om interpretatie te bevorderen kunt U een geschikte *rotatie* van de factorladingenmatrix beschouwen (zie 2.13.). Dit houdt in, dat U i.p.v. de oorspronkelijke factoren andere factoren beschouwt, die lineaire combinaties van de eerste zijn en eveneens in staat zijn de samenhang tussen de variabelen te verklaren. Hierbij is niets wezenlijks toegevoegd of verloren: de oude factoren zijn desgewenst weer terug te vinden uit de nieuwe. Wel dient U de vragen in 1.10. en 1.11. voor de nieuwe, geroteerde ladingenmatrix te beantwoorden! Een voorbeeld van interpretatie vindt U in 2.14.

## 1.13. DATAREDUCTIE

Factoranalyse en principale componentenanalyse kunnen soms worden gebruikt om datareductie te verkrijgen. Daar datareductie niet het onderwerp van dit rapport vormt, zal het hier slechts beknopt worden besproken.

Een datareductiemethode is een techniek om de waarnemingen samen te vatten in een klein aantal nieuwe kenmerken, meestentijds functies van de waarnemingen, zonder dat er veel van de oorspronkelijk in het materiaal aanwezige informatie verloren gaat. De behoefte aan datareductie kan ontstaan, wanneer men, vaak uit angst iets over het hoofd te zien, een (te) groot aantal variabelen per experimentele eenheid heeft gemeten. Door dit grote aantal is het materiaal dan onhandelbaar of er treedt in de verdere analyse storende afhankelijkheid op.

Bij datareductie interesseert de onderzoeker zich minder voor de aard van de samenvattende kenmerken, als ze maar met behoud van informatie zijn materiaal samenvatten. Wat de onderzoeker verstaat onder informatie bepaalt de te gebruiken reductietechniek. Twee (uit vele) mogelijkheden zijn de volgende:

a. Een onderzoeker verricht aan een aantal experimentele eenheden een aantal metingen. Hij ziet dat niet alle meeteenheden dezelfde resultaten geven en hij is nu geïnteresseerd in de oorzaak van de opgetreden verschillen. Daarvoor wil hij weten welke (combinaties van) variabelen de grootste verschillen geven. Een mogelijke technische verwoording hiervan is: zoek combinaties van variabelen die de grootste variantie hebben. Voor deze onderzoeker betekent informatie dus kennis omtrent

verschillen tussen de eenheden en dit wordt uitgedrukt in kennis van de variantie van de variabelen in zijn materiaal. Hij wil een zodanige reductietechniek toepassen dat in het resultaat ook nog die variabelen of combinaties van variabelen met de grootste variantie voorkomen. Hij kan dan principale componentenanalyse gebruiken, daar dit immers een techniek is, die juist díe combinaties van variabelen uit het materiaal selecteert, die zoveel mogelijk van de totale variantie bevatten. (Zie voor toepassing van principale componentenanalyse bijv. MORRISON (1976).

b. Een onderzoeker heeft aan een aantal experimentele eenheden een aantal metingen verricht. Hij wil zijn metingen zodanig reduceren, dat de samenvattende kenmerken het mogelijk maken de samenhang van de originele variabelen te karakteriseren. Een mogelijk technische verwoording hiervan kan geformuleerd worden in termen van correlaties of covarianties. Voor hem betekent informatie dus kennis omtrent de covariantie/correlatiestructuur der variabelen. Hij wil nu die combinaties van variabelen bepalen, die zoveel mogelijk van de oorspronkelijke structuur behouden. Hij kan trachten dit door middel van factoranalyse te verwezenlijken, daar deze techniek de variabelen juist tracht te zien als opgebouwd uit lineaire combinaties van factoren, zodanig dat deze juist de covariantie/correlatiematrix teweeg brengt.

Opgemerkt moet worden, dat datareductie met behoud van variantiestructuur en met behoud van covariantiestructuur beslist niet hetzelfde is (zie 2.8.).

Datareductie is een eenmalige gebeurtenis, die slechts betrekking heeft op het materiaal waarop het wordt uitgevoerd. Er wordt niet gemikt op een uitspraak over een algemenere of toekomstige situatie.

Leest U ook 1.11., waar enige voor datareductie van belang zijnde opmerkingen worden gemaakt.

## 2. UITLEG

De UITLEG bestaat uit een aantal los van elkaar staande onderdelen, waarin een nadere uitweiding over sommige in de LEIDRAAD aangeroerde problemen wordt gegeven. De volgorde van deze secties heeft geen betekenis.

### 2.1. TERMINOLOGIE; DOELEINDEN VAN FACTORANALYSE

Deze sectie dient om een aantal termen, zoals die overal in dit rapport worden gebruikt, in te voeren en om de belangrijkste oogmerken van factoranalyse in deze termen te omschrijven.

We gaan uit van een populatie van experimentele eenheden, waaraan een p-tal kwalitatieve kenmerken (variabelen) kan worden gemeten. Door het aselect trekken van één element uit de populatie, en mogelijk ook door het niet onder controle houden van enige storende invloeden, verkrijgen we de stochastische variabelen $\underline{x}_1,\ldots,\underline{x}_p$ door de kenmerken te observeren; $\underline{x}_1,\ldots,\underline{x}_p$ vatten we samen als een vector $\underline{x} = (\underline{x}_1,\ldots,\underline{x}_p)'$. (Een streepje onder een letter duidt het stochastische karakter aan.) Altijd wordt aangenomen, dat de covariantiematrix van $\underline{x}$, $\Sigma$, bestaat.

De bewering: $\underline{x}$ heeft een *factormodel* (of: $\underline{x}$, of: de verdeling van $\underline{x}$, of: $\Sigma$ vertoont een *factorstructuur*) houdt in: er bestaat een natuurlijk getal $m < p$ en constanten $\mu_i$, $\lambda_{ij}$ $(i = 1,\ldots,p;\ j = 1,\ldots,m)$ en onderling ongecorreleerde stochastische variabelen $\underline{\phi}_j$ $(j = 1,\ldots,m)$ met verwachting 0 en variantie 1, en onderling èn met de $\underline{\phi}_j$ ongecorreleerde stochastische variabelen $\underline{\varepsilon}_i$ $(i = 1,\ldots,p)$ met verwachting 0, zodanig dat

$$\underline{x}_i = \mu_i + \sum_{j=1}^{m} \lambda_{ij}\underline{\phi}_j + \underline{\varepsilon}_i \quad (i = 1,\ldots,p).$$

Of, met de notatie

$$\mu \stackrel{\text{def}}{=} (\mu_1,\ldots,\mu_p)', \qquad \Lambda \stackrel{\text{def}}{=} (\lambda_{ij})_{i=1,j=1}^{p\quad m},$$
$$\underline{\varepsilon} \stackrel{\text{def}}{=} (\underline{\varepsilon}_1,\ldots,\underline{\varepsilon}_p)', \qquad \underline{\phi} \stackrel{\text{def}}{=} (\underline{\phi}_1,\ldots,\underline{\phi}_m)':$$

$$\underline{x} = \mu + \Lambda\underline{\phi} + \underline{\varepsilon}$$

Uitgeschreven voor de k-de experimentele eenheid: (k = 1,...,n)

$$\underline{x}_{ki} = \mu_i + \sum_{j=1}^{m} \lambda_{ij}\underline{\phi}_{kj} + \underline{\varepsilon}_{ki}.$$

De stochastische variabelen $\underline{\phi}_j$ heten *factoren* (in de literatuur soms: gemeenschappelijke factoren). De stochastische variabelen $\underline{\varepsilon}_i$ heten *specifieke gedeelten* (in de literatuur soms: specifieke of unieke factoren). De constanten $\lambda_{ij}$ heten de *factorladingen* van de i-de variabele op de j-de factor. De volgende beweringen gelden, als $\underline{x}$ een factormodel heeft:

$$\lambda_{ij} = \mathrm{cov}(\underline{x}_i, \underline{\phi}_j) \quad (i = 1,...,p;\ j = 1,...,m)$$
$$\mu_i = E\underline{x}_i \quad (i = 1,...,p).$$

De som $\sum_{j=1}^{m} \lambda_{ij}\underline{\phi}_j$ heet het *gemeenschappelijk* of *communale gedeelte* van de i-de variabele $\underline{x}_i$. De grootheid $\mathrm{cov}(\underline{x}_i, \sum_{j=1}^{m} \lambda_{ij}\underline{\phi}_j)$ heet de *communaliteit* van $\underline{x}_i$ en is gelijk aan $\sum_{j=1}^{m} \lambda_{ij}^2$. De grootheid $\mathrm{var}(\underline{\varepsilon}_i)$ heet de *uniciteit* van $\underline{x}_i$ en is gelijk aan $\mathrm{var}(\underline{x}_i)$ minus de communaliteit van $\underline{x}_i$. I.p.v. $\mathrm{var}(\underline{\varepsilon}_i)$ noteren we $\psi_i$.
De bewering "$\underline{x}$ heeft een factormodel met m factoren" is equivalent met de bewering

$$\Sigma = \Lambda\Lambda' + \Psi \quad \text{met} \quad \Psi \stackrel{\text{def}}{=} \begin{bmatrix} \psi_1 & & 0 \\ & \ddots & \\ 0 & & \psi_p \end{bmatrix}.$$

Dit is de vorm waarin een factormodel veelal wordt gepresenteerd en ook de vorm waarvan men uitgaat bij de oplossing ervan. Men noemt een drietal $(m, \Lambda, \Psi)$ een *factoroplossing* voor $\underline{x}$ of $\Sigma$. De geldigheid van een factormodel voor $\underline{x}$ impliceert dat de covariantie tussen elk tweetal $\underline{x}_i$ en $\underline{x}_{i'}$ geheel voortkomt uit de wijze waarop zij van de factoren afhangen, nl.

$$\mathrm{cov}(\underline{x}_i, \underline{x}_{i'}) = \sum_{j=1}^{m} \lambda_{ij}\lambda_{i'j} \quad (1 \le i \ne i' \le p).$$

Daarom zegt men wel dat factoranalyse ten doel heeft de covariantiestructuur van de variabelen te verklaren.

Daar de constanten $\mu_i$ geen rol van betekenis hebben, wordt dikwijls, zonder beperking van algemeenheid, aangenomen dat zij nul zijn, d.w.z. dat de variabelen, waarop de analyse betrekking heeft, gereduceerd zijn. Bij interpretatie van de resultaten moet men daarmee wel rekening houden.

In de literatuur wordt vaak uitgegaan van gestandaardiseerde variabelen, d.w.z. men kiest zodanige schaaltransformaties dat de $\underline{x}_i$ verwachting 0 en variantie 1 hebben. Dit heeft tot gevolg dat men i.p.v. $\Sigma$ de correlatiematrix $\Gamma$ tracht te schrijven als $\Lambda\Lambda' + \Psi$. In dit rapport wordt niet uitsluitend uitgegaan van gestandaardiseerde variabelen.

Het uitvoeren van een factoranalyse kan nu drie verschillende oogmerken hebben,

a. nagaan of $\underline{x}$ een factorstructuur vertoont;
b. nagaan of $\underline{x}$ een factorstructuur met m factoren vertoont (m gespecificeerd);
c. nagaan of $\underline{x}$ een factormodel met gespecificeerde m en $\Lambda$ heeft.

Factoranalyse van het derde type wordt niet behandeld in dit rapport. Zie bijvoorbeeld LAWLEY & MAXWELL (1971). Bij analyses van het eerste of tweede type is men, indien de existentie van een dergelijke factorstructuur niet wordt verworpen, geïnteresseerd in schattingen van m en $\Lambda$ (geval a.) of $\Lambda$ alleen (geval b.).

In het algemeen omvat een factoranalyse

a. aselect trekken van een steekproef van n elementen uit de populatie en het verrichten van de waarnemingen;
b. berekenen van een schatting van de covariantiematrix $\Sigma$;
c. schatten van het aantal factoren;
d. toetsen of een factorstructuur van het vereiste type aanwezig is.

Indien d. bevestigend kan worden beantwoord:

e. schatten van $\Lambda$;
f. interpreteren van de factoren;
g. (eventueel) schatten van de waarde van de factoren bij de elementen uit de steekproef (de *factorscores*).

Dit rapport bespreekt moeilijkheden die bij deze stappen kunnen optreden.

## 2.2. NOTATIE

De volgende symbolen worden door het hele rapport gebruikt, doorgaans zonder verdere toelichting. Overige symbolen worden in de betreffende paragraaf ingevoerd. Een verklaring van de gebruikte termen, voorzover niet bekend verondersteld, is te vinden in 2.1.
Stochastische variabelen worden onderstreept. Met E wordt de verwachting, met $\sigma$ de standaardafwijking van een stochastische variabele genoteerd.

p: aantal variabelen, dimensie van de waar te nemen stochastische variabele

$\underline{x}$: waar te nemen stochastische variabele; $\underline{x} \overset{\text{def}}{=} (\underline{x}_1,\ldots,\underline{x}_p)'$;

m: aantal factoren in een factormodel voor $\underline{x}$;

$\Sigma$: de covariantiematrix van $\underline{x}$;

$\Sigma \overset{\text{def}}{=} E(\underline{x}-E\underline{x})(\underline{x}-E\underline{x})'$
$\Sigma$ is een (p × p)-matrix met op de i,j-de plaats
$\text{cov}(\underline{x}_i,\underline{x}_j) \overset{\text{def}}{=} E\underline{x}_i\underline{x}_j-(E\underline{x}_i)(E\underline{x}_j)$

$\Gamma$: de correlatiematrix van $\underline{x}$; $\Gamma \overset{\text{def}}{=} (\text{diag } \Sigma)^{-\frac{1}{2}} \Sigma(\text{diag } \Sigma)^{-\frac{1}{2}}$; hierin is diag $\Sigma$ een matrix met dezelfde diagonaal als $\Sigma$, doch overigens nullen;
$\Gamma$ is een (p × p)-matrix met op de i,j-de plaats
$\rho(\underline{x}_i,\underline{x}_j) \overset{\text{def}}{=} \text{cov}(\underline{x}_i,\underline{x}_j)/(\sigma(\underline{x}_i)\cdot\sigma(\underline{x}_j))$; waarin $\sigma(\underline{x}_i) = (\Sigma_{ii})^{\frac{1}{2}}$ en $\sigma(\underline{x}_j) = (\Sigma_{jj})^{\frac{1}{2}}$;

$\Lambda$: de (p × m)-matrix der factorladingen;

$\underline{\phi}$: de vector van factoren, $\underline{\phi} \overset{\text{def}}{=} (\underline{\phi}_1,\ldots,\underline{\phi}_m)'$;

$\underline{\varepsilon}$: de vector der specifieke gedeelten, $\underline{\varepsilon} \overset{\text{def}}{=} (\underline{\varepsilon}_1,\ldots,\underline{\varepsilon}_p)'$;

$\Psi$: de covariantiematrix van $\underline{\varepsilon}$; de diagonaalelementen van $\Psi$ worden aangeduid met $\psi_1,\ldots,\psi_p$ en heten uniciteiten (de overige elementen zijn 0);

n: het aantal (p-variate) waarnemingen; aantal onafhankelijke trekkingen van $\underline{x}$ die gedaan worden;

$(\underline{x}_{k1},\ldots,\underline{x}_{kp})$: de k-de onafhankelijke trekking van $\underline{x}$; k = 1,...,n;

$\underline{S}$: $\underline{S} \overset{\text{def}}{=} \left(\frac{1}{n-1} \sum_{k=1}^{n} (\underline{x}_{ki}- \frac{1}{n} \sum_{\ell=1}^{n} \underline{x}_{\ell i})(\underline{x}_{kj}- \frac{1}{n} \sum_{\ell=1}^{n} \underline{x}_{\ell j})\right)_{i=1,j=1}^{p,p}$, d.w.z. de steekproefcovariantiematrix; schatter voor $\Sigma$;

$\underline{C}$: $\underline{C} \overset{\text{def}}{=} (\text{diag } \underline{S})^{-\frac{1}{2}} \underline{S} (\text{diag } \underline{S})^{-\frac{1}{2}}$, d.w.z. de steekproefcorrelatiematrix, schatter voor $\Gamma$;

$\underline{L}$: schatter voor $\Lambda$, te berekenen uit de steekproef;

$\underline{P}$: schatter voor $\Psi$, te berekenen uit de steekproef.

De drie meestgebruikte modellen staan hieronder in formule gespecificeerd. Voor een discussie hiervan zie 2.3.

Het *traditionele of algemene factoranalysemodel* (AFA genoemd in dit rapport) veronderstelt dat $\Sigma$ geschreven kan worden als

$$\Sigma = \Lambda\Lambda' + \Psi \qquad \text{waarin } \Psi \text{ een diagonaalmatrix is.}$$

Andere meer of minder courante namen voor dit model zijn
- Thurstone's multipele factoranalysemodel
- Rao's factoranalysemodel
- Lawley & Maxwell's factoranalysemodel.

Het *gelijke residuele variantiesmodel* (GFA) stelt dat $\Sigma$ geschreven kan worden als

$$\Sigma = \Lambda\Lambda' + \Psi \qquad \text{met } \Psi = \sigma^2 I \text{ voor zekere } \sigma^2 > 0.$$

*Jøreskog's factoranalysemodel* (JFA) stelt dat $\Sigma$ geschreven kan worden als

$$\Sigma = \Lambda\Lambda' + \Psi \qquad \text{met } \Psi = \theta(\text{diag } \Sigma^{-1})^{-1} \text{ voor zekere } \theta \in (0,1),$$

Dit model staat ook bekend als imagefactoranalysemodel (niet te verwarren met imageanalysemodel, zie 2.3).

## 2.3. VERSCHILLENDE MODELLEN EN METHODEN MET HUN VOOR- EN NADELEN

In de literatuur en in de computerprogrammapakketten kan men diverse factoranalysetechnieken aantreffen. Hieraan liggen ten grondslag zowel verschillende factoranalyse*modellen* als verschillende *methoden* om bij een

bepaald model een oplossing te vinden.

De verschillende modellen behelzen verschillende veronderstellingen over de aard van de te analyseren data. Aan deze veronderstellingen moet althans bij benadering zijn voldaan om betrouwbare resultaten te verkrijgen.

Vaak wordt een keuze gemaakt "op praktische gronden": bij voorbeeld hoort men wel eens: "alleen principale componentenanalyse is geschikt, want dat kan als enige draaien met meer variabelen dan waarnemingen".

Een dergelijke formulering doet het onderscheid tussen model en methode al te zeer vervagen. Bij de keuze van een model dient de oplossingsmethode in eerste instantie geen rol te spelen. Als alleen modellen, waarvoor geen oplossingsmethode bestaat, zinvol zijn, is het probleem op grond van de beschikbare waarnemingen niet op te lossen.

Na het kiezen van een model en het vastleggen van veronderstellingen komt het praktische probleem een oplossingsmethode te kiezen uit de verschillende methoden, die bij het gekozen model mogelijk zijn. Voor sommige modellen is er een uniek beste en goed hanteerbare methode. Voor andere modellen kan bij sommige problemen de beste methode te veel rekentijd vergen. In een dergelijk geval moet men het probleem verkleinen (d.w.z. met minder variabelen werken), zo men geen genoegen wil nemen met een minder goede oplossing.

Wij vergelijken nu eerst de mogelijke veronderstellingen en modellen, daarna de mogelijke oplossingsmethoden.

*Modellen*

We beperken ons tot de beschrijving van enkele gangbare modellen. Formele definities zijn vermeld in 3.1.

Allereerst maken wij het belangrijke onderscheid tussen principale componentenanalyse (goedkoop en eenvoudig, maar weinig zeggend in statistische zin) enerzijds, en factoranalyse in eigenlijke zin anderzijds.

A. *Principale componentenanalyse* (PCA)

Principale componentenanalyse tracht de positie van de multivariate waarnemingen in de puntenwolk zo zuinig mogelijk weer te geven, en wel

door zo weinig mogelijk lineaire combinaties van de oorspronkelijke variabelen.

Principale componentenanalyse is in feite een rekentechniek, waaraan geen passend stochastisch model ten grondslag ligt. Met of zonder rotatie van belangrijke componenten levert het een mogelijke, maar zeker niet de enige, aanpak van het datareductieprobleem. De resultaten kunnen alleen naar subjectieve maatstaven worden beoordeeld op grond van het succes van de verdere analyse. PCA geeft geen oplossing voor het factoranalysemodel, zie 2.9, terwijl de schaalafhankelijkheid van PCA vaak een probleem is, zie 2.10.

B. *Factoranalyse in eigenlijke zin*

Factoranalysemodellen veronderstellen algemeen dat iedere variabele de som is van een gemeenschappelijk of communaal gedeelte en een specifiek of uniek gedeelte; dat de communale gedeelten van de variabelen samenhang met elkaar vertonen en dat de unieke gedeelten geen samenhang met elkaar noch met de communale gedeelten bezitten.

Zonder meer is deze stelling volkomen leeg: er is altijd aan voldaan en wel op vele wezenlijk verschillende manieren. Meer veronderstellingen, bij voorbeeld over de aard van de gemeenschappelijke delen zijn nodig voordat het gepostuleerde model zinvol wordt.

De meest gangbare extra veronderstellingen zijn die van het algemene model.

B.1. *Algemene of traditionele factoranalysemodel* (AFA)

Het algemene model voegt aan bovenstaande de veronderstelling toe, dat de gemeenschappelijke gedeelten lineaire functies zijn van een m-tal *factoren* $\underline{\phi}_1,\ldots,\underline{\phi}_m$. Dit is het model dat in 2.1 in extenso is beschreven, culminerend in de klassieke formule $\Sigma = \Lambda\Lambda' + \Psi$.

Wanneer men zinvolle en betrouwbare resultaten tracht te behalen, is het nodig te veronderstellen dat het gepostuleerde factormodel uniek is, d.w.z. dat geen andere factorladingenmatrix $\Lambda$, uniciteiten $\Psi$, factoren $\underline{\phi}_1,\ldots,\underline{\phi}_m$ en specifieke gedeelten $\underline{\varepsilon}_1,\ldots,\underline{\varepsilon}_p$ bestaan die evenzeer bij de observaties passen.

Het is spijtig dat dit zeer populaire model nu juist lijdt aan een hele serie intrinsieke gebreken wat betreft uniekheid.

In de eerste plaats kan het voorkomen, dat er bij vaste m verschillende $\Psi$'s bestaan zodanig dat $\Sigma$ ontbonden kan worden met die $\Psi$. Verschillende oplossingsmethoden zijn niet bestand tegen een dergelijke situatie. Een kant en klare remedie tegen dit zogenaamde *modelidentificatieprobleem* is niet bekend. Wel moet, om dit te voorkomen, zeker gelden $m < \frac{1}{2}(2p + 1 - \sqrt{8p + 1})$. Voor verdere discussies verwijzen wij naar 2.6.

In de tweede plaats zijn er bij gegeven m en $\psi$ altijd oneindig veel ladingenmatrices $\Lambda$. Dit is het zogeheten *rotatieprobleem* waaraan gedeeltelijk tegemoet gekomen kan worden, in verband met de interpretatie (zie 2.13).

Hieraan is veel meer aandacht besteed dan aan het derde probleem, inhoudende, dat gegeven m, $\Lambda$ en $\Psi$ er nog oneindig veel, vaak wezenlijk verschillende $\underline{\phi}_j$ en $\underline{\varepsilon}_i$ bestaan, die aan het model voldoen. Dit zogenaamde *factoridentificatieprobleem* is een serieus struikelblok voor degenen die een interpretatie -een betekenis- aan de gevonden factoren willen toekennen, of factorscores willen uitrekenen voor elke experimentele eenheid. Zoals gezegd bestaat voor dit probleem geen oplossing, wel kan men aan de hand van het *Guttman criterium* beoordelen of het in een gegeven geval een ernstig probleem is. Zie 2.12.

In het algemeen kan men zeggen dat aan het algemene model te gemakkelijk wordt voldaan. Het model van Jøreskog en het gelijke residuele variantiesmodel, die hierna besproken zullen worden, leggen dan ook extra voorwaarden op, waardoor het modelidentificatieprobleem wordt voorkomen (niet echter het factoridentificatieprobleem). Deze voorwaarden zijn echter nogal streng of gekunsteld.

Wanneer men aan de veronderstellingen van het algemene model nog kan toevoegen dat de variabelen multinormaal verdeeld zijn, verkrijgt men grote voordelen i.v.m. de evaluatie van de gevonden oplossing, vgl. 1.8, 1.9 en 1.10.

### B.2. *Het gelijke residuele varianties factoranalysemodel* (GFA)

Het gelijke residuele variantiesmodel is het algemene model met een

extra eis opgelegd aan de uniciteiten (die soms ook residuele varianties worden genoemd, vandaar de naam), n.l. $\psi_1 = \psi_2 = \ldots \psi_p$. Om deze veronderstelling een schijn van redelijkheid te geven, zal er een zekere symmetrie of verwisselbaarheid tussen de variabelen moeten bestaan en ze zullen in een zekere zin natuurlijke, uitverkoren schaal gemeten moeten zijn, daar een GFA-model schaaltransformaties, die niet gelijk zijn voor alle variabelen, niet weerstaat (zie 2.10).

Door de zwaarte van deze eisen is het GFA-model nooit erg populair geworden, maar als deze veronderstellingen vervuld zijn is het een zeer handzaam model, omdat t.o.v. AFA het modelidentificatieprobleem (niet het factoridentificatieprobleem) uit de wereld is geholpen. De schatting van het model verloopt ook erg eenvoudig.

B.3. *Jøreskog's factoranalysemodel* (JFA)

Jøreskog, niet tevreden met de extra eisen van het GFA-model, stelde een andere aanvullende eis t.o.v. AFA, die eveneens het modelidentificatieprobleem oplost en voor de oplossing voordelen biedt. Jøreskog's eis luidt: $\Psi$ is evenredig met $(\text{diag}\ (\Sigma^{-1}))^{-1}$.
Het voordeel van deze voorwaarde boven de GFA-voorwaarde is dat JFA wel schaaltransformaties weerstaat. Maar met welke redenen zou men ooit durven uitspreken of aan deze eis voldaan is?

JØRESKOG (1963) geeft zelf de volgende (theoretische) rechtvaardiging: Stel dat een oneindige rij variabelen $\underline{x}_1, \underline{x}_2, \ldots$ is waargenomen en dat de i-de variabele $\underline{x}_i$ gemeenschappelijk gedeelte $\underline{g}_i$ heeft en specifiek gedeelte $\underline{\varepsilon}_i = \underline{x}_i - \underline{g}_i$. Stel dat AFA geldt voor $(\underline{x}_1, \ldots, \underline{x}_p)$ voor elke p, met een zeker eindig aantal factoren m. Dan geldt ook JFA asymptotisch dan en slechts dan als de $\underline{g}_i$ "asymptotisch" identificeerbaar zijn (en dan ook uit $\underline{x}_1, \underline{x}_2, \ldots$ te berekenen). In dit geval naderen de waarden van het Guttman criterium asymptotisch tot 1.

Jøreskog redeneert dat we alleen wat aan factoranalyse hebben als we in staat zijn om, tenminste asymptotisch, de factoren, als objecten van onze studie, te determineren, d.w.z. als dus de $\underline{g}_i$ identificeerbaar zijn. In die situatie is dus ten naaste bij aan JFA voldaan. Deze rechtvaardiging van Jøreskog's bijvoorwaarde garandeert echter allerminst, dat deze ook vervuld is.

B.4. *Andere varianten van het algemene model*

Vele variaties op AFA zijn nog denkbaar en sommige zijn ook werkelijk gerealiseerd: bijvoorbeeld nadere specificatie in het model van $\Lambda$, zoals waar nullen en niet-nullen moeten staan of het voorschrijven van de grootte van bepaalde elementen. Dit kan soms ook het modelidentificatieprobleem oplossen en veelal wordt het rotatieprobleem ermee omzeild. In dit rapport worden deze modificaties niet behandeld.

B.5. *Imageanalyse*

Aan imageanalyse ligt een model ten grondslag dat, hoewel misschien van psychometrische betekenis, nog geen bevredigende wiskundige gedaante heeft.

De ontwerper ervan, GUTTMAN (1953), gaat er van uit dat de variabelen $\underline{x}_1,\ldots,\underline{x}_p$ een greep zijn uit een universum van alle mogelijke relevante variabelen; verder definieert hij het gemeenschappelijk gedeelte van een variabele als de beste predictor van die variabele op grond van alle andere variabelen in het universum. Er zijn verbanden met Jøreskog's model, dat Jøreskog zelf niet voor niets "imagefactoranalysemodel" noemde (JØRESKOG (1969)), een naam die wij terwille van de duidelijkheid vermijden.

Het factoridentificatieprobleem is hier formeel opgelost door factoren te definiëren als functie van waarneembare variabelen, maar men heeft weinig aan deze definitie zonder een nadere uitwerking van wat "alle mogelijke relevante variabelen" eigenlijk zijn en van wat de beste schatter van een variabele gebaseerd op oneindig veel andere precies inhoudt.
Aan imageanalyse wordt in dit rapport verder geen aandacht besteed (behoudens in 3.8) omdat er geen statistisch model aan ten grondslag ligt.

B.6. *Alpha-analyse*

Ook alpha-analyse maakt gebruik van het psychometrische concept van "alle relevante variabelen". Factoren zijn gedefinieerd als variabelen die een positieve generaliseerbaarheid in termen van Cronbach's coëfficiënt $\alpha$ (LORD & NOVICK (1968)) hebben. Het model is ontwikkeld door KAISER & CAFFREY (1965) en wordt hier verder niet beschouwd (behoudens in 3.8), omdat er geen statistisch model aan ten grondslag ligt.

*Methoden*

Voor sommige modellen zijn verschillende alternatieve oplossingsmethoden voorgesteld, voor andere bestaat overeenstemming over wat men het beste kan doen. We bespreken nu per model de oplossingsmethoden, waarbij ook aandacht besteed wordt aan beschikbaarheid van programmatuur, zonder dat dit een oordeel inhoudt over de kwaliteiten van die programmatuur.

A. Voor PCA bestaat een eenvoudige en directe methode van oplossen door het berekenen van eigenwaarden en eigenvectoren van $\Sigma$, of $\Gamma$. Zie bijvoorbeeld MORRISON (1976). De methode is veelvuldig geprogrammeerd, o.a. in STATAL (BETHLEHEM (1976)) en in SPSS (NIE e.a. (1975)). In het laatstgenoemde pakket wordt hij ten onrechte als een soort factoranalyse gepresenteerd.

B.1. Voor AFA zijn talloze methoden voorgesteld, waarvan wij hier bespreken:
   a. de directe principale componentenmethode, DPC
   b. de indirecte principale componentenmethode, IPC
   c. de maximum likelihoodmethoden volgens Lawley & Maxwell of Rao, MLR
   d. de maximum likelihoodmethode volgens Jøreskog, MLJ.

   Over de benaming van de methoden bestaat geen overeenstemming en het is vaak moeilijk na te gaan welke methode in een concreet geval toegepast is. Bovenstaande benamingen zijn van ons.

   We wijzen erop, dat aan "Rao's canonieke factoranalyse" (RAO (1955)) hetzelfde model ten grondslag ligt als aan AFA. We spreken daarom niet over een apart Rao-model, wel over de oplossingsmethode volgens Rao, die in hoofdzaak overeenkomst met die van Lawley en Maxwell, samengevat als MLR. Ook is het verwarrend dat Jøreskog zowel een apart model, JFA, heeft bedacht, alsook een oplossingsmethode MLJ voor het algemene model; maar zo liggen de zaken nu eenmaal.

   Er zijn nog wel andere methoden in omloop om AFA uit te voeren, zoals bijvoorbeeld de centroïdmethode. Deze zijn echter achterhaald door bovengenoemde methoden, die wel alle gebruik maken van een rekenautomaat.

   Recent hebben VAN DRIEL e.a. (1974) een methode voorgesteld, die speciaal geschikt is voor zogeheten oneigenlijke oplossingen, zie 3.6. Deze methode ziet er veelbelovend uit, maar hij is nog niet gemakkelijk algemeen beschikbaar.

*Directe principale componentenmethode* (DPC)

Door sommigen wordt aanbevolen de oplossingsmethode van principale componentenanalyse als oplossingsmethode voor AFA te gebruiken. (Deze schrijvers maken vaak ook geen onderscheid tussen PCA en factoranalyse.)

Zelfs als het aantal factoren waarin $\Sigma$ exact kan worden ontbonden bekend is, vindt deze methode in het algemeen geen oplossing die aan het AFA-model voldoet. De methode moet dan ook worden ontraden. De methode is geprogrammeerd in SPSS onder de naam PC1 (NIE e.a. (1975)).

*Indirecte principale componentenmethode* (IPC)

Met deze naam bedoelen wij het iteratieve algorithme zoals beschreven staat bij HARMAN (1960), p. 135 e.v.. Sommige schrijvers spreken van de principale factorenmethode. Dit is wel de meest populaire oplossingsmethode. Van zijn eigenschappen is theoretisch weinig bekend, er zijn gevallen bekend waar de methode de correcte oplossing niet vindt, maar ook vele, waar het redelijk lijkt (vgl. 3.4, 3.8). De methode geeft veelvuldig aanleiding tot oneigenlijke oplossingen (zie 3.6) en eist soms erg veel iteratieslagen, voordat een redelijk stabiel resultaat is bereikt. We raden de methode niet erg aan. De methode is geprogrammeerd in SPSS onder de naam PC2 (NIE e.a. (1975)).

De beide andere methoden, MLR en MLJ, zijn gebaseerd op de berekening van maximum likelihoodschatters voor $\Lambda$ en $\Psi$ onder aanname van multinormale verdeling van de waarnemingen. Bewezen kan worden (zie 3.7) dat deze schatters ook zonder normaliteit nog een aardige eigenschap bezitten, nl. die van asymptotische raakheid, d.w.z. dat men met elke gewenste graad van zekerheid elke gewenste precisie van de schatters kan verkrijgen door maar een groot genoeg aantal waarnemingen te doen.

*Maximum likelihoodmethode volgens Lawley & Maxwell of Rao* (MLR)

De MLR-methoden om deze maximum likelihoodschatters te berekenen deugen helaas niet, in die zin, dat het iteratieve algorithme in de praktijk niet binnen een redelijk aantal iteratieslagen tot stabiele uitkomsten leidt. Ook is niet bewezen dat de algorithmen op den duur wel tot conver-

gentie leiden. De methode van Rao is geprogrammeerd in SPSS onder de naam RAO (NIE e.a. (1975)).

*Maximum likelihoodmethode volgens Jøreskog* (MLJ)

De MLJ-methode is zonder twijfel thans de meest geschikte om het AFA-model op te lossen. De methode bestaat eveneens uit een iteratief algorithme, doch dit leidt binnen een klein tot zeer klein aantal interatieslagen tot bevredigende resultaten. MLJ gebruikt de JFA-oplossing als eerste iteratieslag. Theoretisch is bewezen dat de methode altijd convergeert. Voor grote aantallen variabelen, zeg meer dan 30, is de methode nogal tijdrovend.

De methode is door Jøreskog geprogrammeerd. Een versie van dit programma is beschikbaar in STATAL.

B.2. Voor GFA bestaat een eenvoudige, niet-iteratieve en daardoor goedkope methode, beschreven in 3.1, die de maximum likelihoodschatters onder multinormaliteit geeft.
De methode is geprogrammeerd in STATAL.

B.3. Voor JFA heeft JØRESKOG (1963) een eenvoudige, niet-iteratieve en daardoor goedkope methode aangegeven. De door deze methode gegeven schatting is een benadering voor de maximum likelihoodschatting voor het JFA-model onder multinormaliteit.
De methode is geprogrammeerd in STATAL.

Verder heeft JØRESKOG (1969) een methode geconstrueerd om de maximum likelihoodschatters voor het JFA-model te schatten. De methode is wederom iteratief en heeft eigenschappen, vergelijkbaar met die van MLJ. Het schijnt ons evenwel toe dat er zelden reden is om deze echte maximum likelihoodschatters te prefereren boven de niet-iteratieve JFA-oplossing. De methode is, voorzover ons bekend, niet in een programmapakket beschikbaar.

De lezer wordt overigens verzocht goed uit elkaar te houden:
a. Jøreskog's factoranalysemodel, JFA
b. Jøreskog's oplossingsmethode voor het algemene model, MLJ

c. Jøreskog's oplossingsmethode voor Jøreskog's model, zoals hierboven besproken.

B.4./B.5. Voor de volledigheid vermelden wij dat voor imageanalyse en alpha-analyse door hun ontwerpers ook oplossingsmethoden zijn ontworpen (GUTTMAN (1953), KAISER & CAFFREY (1965)). Deze methoden zijn geprogrammeerd in SPSS (NIE e.a. (1975)).

Veel programmatuur specificeert op nogal twijfelachtige wijze een aantal factoren als de gebruiker dat niet doet (vgl. 2.10 en 3.8) en zelden wordt gebruik gemaakt van een modeltoets, schatting van de nauwkeurigheid van de schatters en het Guttman criterium. In STATAL is getracht hierin verbetering aan te brengen.

Samenvattend kunnen we geen definitieve aanbeveling doen voor de modelkeuze. Met het oog op bovenstaande discussie en 2.4 en 3.8 kunnen we wel het volgende zeggen: In de eerste plaats beperken we ons tot het AFA-, JFA- en GFA-model met de bijbehorende aanbevolen oplossingsmethode. Dit zijn immers statistische modellen, waarvan de juistheid (tenminste deels) objectief onderzocht kan worden en waarvan de oplossingsmethoden prettige statistische eigenschappen hebben.

JFA en GFA verschillen daarin van AFA dat ze beperkingen zijn van AFA. Deze beperkingen zijn zodanig dat het identificatieprobleem, wat betreft het aantal factoren m en de covariantiematrix $\Psi$ van de specifieke variabelen, opgelost is en de oplossingsmethode sterk vereenvoudigd wordt. Hiervoor moeten echter dan wel veronderstellingen omtrent het model gedaan worden die in het algemeen niet hoeven te gelden (ook niet bij benadering). Hoewel dit in het schema niet is verwerkt lijkt het ons daarom raadzaam om in ieder geval AFA uit te voeren. Zijn er aanwijzingen dat het JFA- of GFA-model zou kunnen gelden (wegens de opmerking op pagina 26 bij B.3 geldt JFA bij benadering als alle Guttman criteriumwaarden zeer hoog zijn), dan kan men deze modellen proberen.

Leidt het gebruik van de drie modellen echter tot duidelijk verschillende uitkomsten, dan verdient het toch aanbeveling om zich eerst te bezin-

nen over de oorzaak hiervan alvorens conclusies te trekken over de theorie die men onderzoekt.

In AFA kan men de mate van identificeerbaarheid van $\Psi$, gegeven m, in een klein probleem onderzoeken door bestudering van de aannemelijkheidsfunctie van $\Psi$. Wanneer deze functie in de buurt van zijn maximum niet veel varieert, of wanneer er meerdere maxima zijn, kan dit een indicatie zijn voor onidentificeerbaarheid. Deze maxima kunnen mogelijk gevonden worden door voor verschillende startwaarden van de schatting van $\Psi$ de maximalisatie van de aannemelijkheidsfunctie uit te voeren.

## 2.4. RESTRICTIVITEIT EN IDENTIFICEERBAARHEID VAN FACTORMODELLEN

Er zijn twee redenen waarom een factoroplossing voor een zeker probleem nietszeggend kan zijn:
a. omdat het factormodel niet restrictief is
b. omdat het factormodel niet identificeerbaar is.

*Restrictiviteit*

Wanneer men structuur in multivariaat waarnemingsmateriaal tracht te ontdekken hoopt men empirisch relevante, en in ieder geval falsifieerbare, uitspraken te kunnen doen. Wat houdt dit in?

Vergelijk het volgende voorbeeld:
Iemand doet onderzoek naar het aantal Surinamers in Nederland dat wel zou willen repatriëren en hij publiceert de uitspraak: dit aantal is ontbindbaar in priemfactoren. Hij heeft dan niet bijgedragen aan de kennis omtrent Surinamers. Immers, elk aantal is ontbindbaar in priemfactoren. Dit wiskundige feit legt geen restricties op aan aantallen: er zijn geen mogelijke aantallen die deze stelling zouden kunnen falsificeren. De uitspraak is een tautologie en zegt derhalve niets over aantallen Surinamers. Had de onderzoeker daarentegen gepubliceerd: dit aantal is deelbaar door 100, dan had hij wel een restrictieve uitspraak gedaan omdat hierdoor een hoeveelheid aantallen, zoals bijvoorbeeld 791, is uitgesloten. Een dergelijke uitspraak wordt *restrictief* genoemd omdat hij de mogelijkheden beperkt, of *falsificeerbaar* omdat het aantreffen van een volgens deze uitspraak onmogelijke stand van zaken de onjuistheid van de uitspraak aantoont, en *empirisch relevant*

omdat alleen empirisch onderzoek (i.t.t. logische analyse) de uitspraak kan falsificeren.

Merk op dat empirische relevantie niet inhoudt, dat de uitspraak ook belangrijk of interessant is. In bovenstaand voorbeeld blijkt dat maar al te duidelijk. Doch een uitspraak kan slechts belangrijk zijn als hij empirisch relevant is; dit is dus eigenlijk een minimumeis.

We zullen nu verder steeds de term restrictief uit bovenstaand drietal gebruiken. In factoranalytisch onderzoek wil men tot uitspraak komen als:

a. deze p-dimensionale stochastische variabele heeft een factormodel óf
b. deze p-dimensionale variabele heeft een m-factormodel (met zekere gespecificeerde m).

Uitspraak a. is niet restrictief. Voor elk van de drie modellen (AFA, JFA, GFA) geldt: voor elke p-variate $\underline{x}$ bestaat een (p - 1)-factoroplossing.

Dit impliceert tevens dat b. voor m = p - 1 ook niet restrictief is. De extra eisen die in JFA en GFA aan de residuele varianties worden opgelegd maken b. restrictief voor elke waarde van $m < p - 1$. Ieder model legt essentieel restricties op, ook het algemene model, hoewel daar de zaak wat moeilijker ligt. In feite is in AFA niet zonder meer duidelijk voor welke waarden van m en p uitspraak b. restrictief is.

Voor m = p - 1 is het algemene model in ieder geval niet restrictief, daar er dan voor willekeurige $\Sigma$ altijd een ontbinding $\Sigma = \Lambda\Lambda' + \Psi$ bestaat. Er bestaat een natuurlijk getal m' zodat AFA wel restrictief is voor $m < m'$ en niet voor $m \geq m'$. Voor $m < m'$ kunnen alleen bijzondere $\Sigma$'s ontbonden worden, voor $m \geq m'$ kunnen alleen bijzondere $\Sigma$'s niet ontbonden worden. In het laatste geval is het geen interessante uitspraak om te zeggen dat $\Sigma$ ontbonden kan worden, omdat dit voor bijna alle $\Sigma$ mogelijk is. In het eerste geval kunnen we zeggen dat het algemene model restrictief is. De waarde van m' wordt als volgt afgeleid:

Zij $\Sigma$ de covariantiematrix van $\underline{x}$ en noteer een m-factoroplossing voor $\Sigma$ met $(m, \Lambda, \Psi)$, en beschouw een vaste rotatiemethode voor $\Lambda$ (vgl. 2.13), noteer de geroteerde matrix met R. Voor welke m is de uitspraak:

"$\Sigma$ heeft een factoroplossing $(m, R, \Psi)$" restrictief?

Beschouw daartoe de gelijkheid

$$\Sigma = RR' + \Psi$$

als een stelsel van $\frac{1}{2}p(p + 1)$ vergelijkingen in $pm + p$ onbekenden (de elementen van R en de diagonaalelementen van $\Psi$). Het spreekt vanzelf dat we niet geïnteresseerd zijn in complexe oplossingen. Het stelsel moet daarnaast voldoen aan $\frac{1}{2}(m - 1)m$ extra condities, welke voortvloeien uit de speciale keuze van de rotatiemethode. Als $\Lambda$ een oplossing is, kan deze zodanig geroteerd worden, dat hij overgaat in een onderdiagonaalmatrix, d.w.z. een matrix $\Lambda$ met $\lambda_{ij} = 0$ voor $j > i$. Dit zijn er in totaal $\frac{1}{2}(m - 1)m$. Derhalve zijn er geen $pm + p$, maar $pm + p - \frac{1}{2}(m - 1)m$ vrije parameters. Wanneer we nu even buiten beschouwing laten dat bovendien voldaan moet zijn aan de ongelijkheden $0 < \psi_i < \sigma_{ii}$, met $\psi_i$ het i-de diagonaalelement van $\Psi$ en $\sigma_{ii}$ dat van $\Sigma$, $i = 1,\ldots,p$, en dat alle oplossingen reëelwaardig moeten zijn, dan volgt dat in het algemeen een oplossing voor *elke* $\Sigma$ mogelijk is, wanneer het aantal vergelijkingen plus het aantal condities kleiner is dan of gelijk is aan het aantal onbekenden. In dat geval is het algemene factormodel dus niet restrictief en we zijn derhalve slechts geïnteresseerd in oplossingen waarin het aantal vergelijkingen en condities groter is dan het aantal onbekenden, dus wanneer geldt

$$\frac{1}{2}p(p + 1) + \frac{1}{2}(m - 1)m > pm + p.$$

Enige manipulaties herleidt dit tot:
Het algemene model is restrictief wanneer geldt:

$$m < \frac{1}{2}(2p + 1) - \sqrt{8p + 1}\,.$$

In verband met het voorbehoud t.a.v. de ongelijkheden voor $\psi_i$ en de reëelwaardigheid van de oplossing is het omgekeerde niet juist: er bestaan covariantiematrices die voor geen enkele waarde van m ($< p - 1$) een factoroplossing hebben (bijv. $\Gamma_3$ in 3.4). Hoe één en ander precies zit wanneer bovenstaande ongelijkheid niet geldt, is onbekend. Voor de zekerheid bevelen wij deze grens aan.

Dat de niet-restrictiviteit een wezenlijk probleem is van het factormodel moge blijken uit het volgende. We veronderstellen dat p variabelen een m-factorstructuur hebben, maar dat we slechts een geringer aantal ($p^*$) variabelen hebben kunnen waarnemen. m is niet restrictief bij de $p^*$ variabelen, zodat we bij de analyse een te kleine waarde $m^*$ van het aantal factoren kunnen vinden, waardoor dan $\Psi$ en $\Lambda$ verkeerd berekend worden.

*Identificeerbaarheid*

Gegeven een (p×p)-covariantiematrix $\Sigma$, verstaan we onder factoroplossing van $\Sigma$ een drietal $(m, \Lambda, \Psi)$ zodanig dat $\Sigma = \Lambda\Lambda' + \Psi$. De vraag naar identificeerbaarheid van een model is de vraag of er, zo er een oplossing van het model is, ook meer precies één oplossing bestaat, of althans slechts op niet essentiële wijze van elkaar verschillende oplossingen.

Als er meerdere essentieel verschillende oplossingen bestaan is het niet duidelijk hoe wij de resultaten van een analyse moeten interpreteren. Daarnaast is er het gevaar dat vooral iteratieve oplossingsmethoden van matige kwaliteit (vgl. 2.3) een "oplossing" zullen verschaffen die tussen verscheidene goede oplossingen inligt en dus de plank misslaan. Ook weet men vaak niet of er nog meer oplossingen bestaan, zodat men niet zeker kan zijn of een eventuele werkelijke factorstructuur ook gevonden is.

Daarom is het zinvol de identificeerbaarheid van de verschillende factoranalysemodellen te bestuderen. Daarbij beperken we ons expliciet tot de studie van de identificeerbaarheid van *restrictieve* modellen.

Nu is geen der drie modellen identificeerbaar, maar we zullen een iets gedetailleerder studie maken van de aard van die niet-identificeerbaarheid. Een kenmerk van een factormodel heet identificeerbaar als voor elk tweetal oplossingen dit kenmerk overeenstemt. (Merk op dat dit niet inhoudt dat het ook een eenvoudige zaak is achter dit kenmerk te komen.)

Wij onderscheiden nu de volgende kwesties:

a. Is het aantal factoren, m, identificeerbaar?
b. Is, gegeven m, de uniciteitenmatrix $\Psi$, identificeerbaar?
c. Is, gegeven m en $\Psi$, de ladingenmatrix $\Lambda$, identificeerbaar?
d. Zijn, gegeven m, $\Psi$ en $\Lambda$, de factoren $\underline{\phi}_1, \ldots, \underline{\phi}_m$ identificeerbaar?

De vierde van deze vragen krijgt een aparte behandeling in 2.12, de overige worden hier behandeld.

*Identificeerbaarheid van het aantal factoren*

Wij beschouwen eerst het algemene model AFA. REIERSØL (1950) bewees hiervoor: voor elke $\Sigma$ impliceert het bestaan van een m-factoroplossing het bestaan van oneindig veel oplossingen $(m + 1, \Lambda_{m+1}, \Psi)$ met verschillende uniciteitenmatrices $\Psi$.

Dit impliceert dat, zo $\Sigma$ een AFA-factormodel heeft, hij altijd meerdere factoroplossingen met verschillende m toelaat, dus is het aantal factoren in AFA niet identificeerbaar.

Voor Jøreskog's model JFA en het gelijke residuele variantiesmodel GFA geldt dat het aantal factoren wel identificeerbaar is (JØRESKOG (1963), zie 3.1). Evenwel geldt wel, dat wanneer een m-factor JFA- of GFA-model geldt, eveneens oneindig veel (m + 1)-factor AFA-modellen gelden, die dan echter géén JFA- (GFA-)modellen zijn. Dit volgt uit Reiersøl's stelling, daar elk JFA- (GFA-)model een bijzonder AFA-model is.

In strikte zin is dit geen niet-identificeerbaarheid omdat het oplossingen voor verschillende modellen betreft, maar de gevolgen voor interpretatie zijn gelijksoortig: het is niet duidelijk aan welke mathematische mogelijkheid wij de voorkeur moeten geven.

Wel identificeerbaar bij AFA is het kleinste aantal factoren waarvoor een factoroplossing mogelijk is bij $\Sigma$. Dit aantal heet de *minimale rang* $m_0$ van $\Sigma$. Op het feit dat de minimale rang $m_0$ identificeerbaar is, berust het gebruik om door successief proberen met een opklimmend aantal factoren een oplossing te zoeken.

*Identificeerbaarheid van de uniciteitenmatrix bij een gegeven aantal factoren*

Voor het algemene model is het antwoord op de vraag: "is $\Psi$ identificeerbaar, gegeven m?" niet voor elke covariantiematrix $\Sigma$ gelijk.

WILSON & WORCESTER (1939) hebben een aantal voorbeelden gegeven van

covariantiematrices $\Sigma_1$, $\Sigma_2$ en $\Sigma_3$, die sprekend op elkaar lijken, waarvan de eerste matrix twee oplossingen voor $\Psi$, de tweede één, en de derde geen oplossing toelaat bij de minimale rang $m_0$ (zie 3.4). Voor $\Sigma_1$ en $\Sigma_2$ is $m_0 = 3$ en voor $\Sigma_3$ is dus $m_0 > 3$. De twee oplossingen die $\Sigma_1$ toelaat hebben daarbij sterk verschillende uniciteitenmatrix.

Deze situatie is zeer onaangenaam: verschillende $\Psi$ kunnen, via de eruit voortvloeiende verschillende $\Lambda$, zeer verschillende interpretatie toelaten, maar men kan aan zijn materiaal niet zien, of $\Sigma$ meerdere ontbindingen toelaat! (Slechts indien men $\Sigma$ exact kent en niet slechts over een schatting ervan beschikt, kan men trachten om enige partiële resultaten van ANDERSON & RUBIN (1956) te gebruiken. Een van die resultaten zegt dat het meestal wel goed is als $m < \frac{1}{2}p$.)

Deze onaangename stand van zaken bij het algemene model is in feite de reden geweest voor de introductie van Jøreskog's model en het gelijke residuele variantiesmodel, die beide voldoende extra eisen opleggen om $\Psi$ wel identificeerbaar te maken, gegeven m. Over de redelijkheid van die eisen valt te twisten (zie 2.3), maar identificeerbaarheid van $\Psi$ leveren ze wel op.

*Identificeerbaarheid van de ladingenmatrix $\Lambda$, gegeven m en $\Psi$*

Wanneer $(m, \Lambda, \Psi)$ een oplossing voor een factormodel voor $\Sigma$ is, en $\theta$ is een orthogonale $(m \times m)$-matrix, dan is $(m, \Lambda\theta, \Psi)$ ook een oplossing. (Een orthogonale matrix $\theta$ is een vierkante matrix die voldoet aan $\theta\theta' = I$.) Immers geldt:

$$\Sigma = \Lambda\Lambda' + \Psi = \Lambda I\Lambda' + \Psi = \Lambda\theta\theta'\Lambda' + \Psi = (\Lambda\theta)(\Lambda\theta)' + \Psi.$$

Dit impliceert dat de ladingenmatrix niet identificeerbaar is, bij gegeven m en $\Psi$, maar men tracht van de nood een deugd te maken door, zoekend naar een interpretatie, een aannemelijke rotatie te vinden. Men neemt zich voor uit alle $\theta$ diegene te kiezen die $\Lambda\theta$ aan zekere extra eisen laat voldoen die geacht worden de interpretatie aan waarde te laten winnen (zie 2.13). Men zou kunnen zeggen, dat interpreteerbaarheid een soort identificeerbaarheid levert.

Wanneer men een zekere vaste rotatiemethode kiest, zeg zodanig i.p.v. een gevonden oplossing $\Lambda$ altijd de geroteerde oplossing $R(\Lambda)$ wordt opgegeven, dan is, gegeven $\Psi$ en m $R(\Lambda)$ identificeerbaar.

Bovenstaande geldt voor elk der drie beschouwde modellen.

Over de vierde vraag, naar identificeerbaarheid van $\underline{\phi}_1,\ldots,\underline{\phi}_m$ bij gegeven m, $\Lambda$, $\Psi$ handelt 2.13. Hier vermelden we slechts dat hier altijd niet-identificeerbaarheid bestaat, maar in verschillende mate. Samenvattend gelden de volgende condities voor restrictiviteit en identificeerbaarheid.

| | Restrictiviteit | Identificeerbaarheid | | |
|---|---|---|---|---|
| | | m | $\Psi \mid m$ | $\Lambda \mid \Psi, m$ |
| Algemene model | $m < \frac{1}{2}(2p + 1 - \sqrt{8p + 1})$ vuistregel | $m_0$ | onzeker | op rotatie na |
| Jøreskog's model | $m < p - 1$ | mits $m < p - 1$ | mits $m < p - 1$ | op rotatie na |
| Gelijke residuele variantiesmodel | $m < p - 1$ | mits $m < p - 1$ | mits $m < p - 1$ | op rotatie na |

In de overige gedeelten van dit rapport wordt, al dan niet stilzwijgend, aangenomen, dat het om restrictieve en identificeerbare factormodellen en oplossingen gaat. I.h.b. houdt dit voor het algemene model in, dat het om de minimum rangoplossing gaat.

## 2.5. STOCHASTIEK IN FACTORMODELLEN

Wanneer men uitspraken wil doen over een groter geheel dan het waargenomen materiaal en men wil daarbij van de machinerie van de wiskundige statistiek gebruikmaken, dan is het noodzakelijk dat het redelijk is de waarnemingen voor te stellen als realisaties van een of meer stochastische variabelen. Wanneer dit een redelijke aanname is en wanneer niet, is niet

gemakkelijk te zeggen. In feite is over deze zeer fundamentele vraag verbazingwekkend weinig nagedacht, misschien omdat het zo moeilijk is er in het algemeen iets over te zeggen.

In een factoranalytische context zijn de volgende overwegingen wellicht nuttig, waarmee niet gezegd wil zijn dat de hier omschreven manier de enige zinnige is om een situatie van passende stochastiek te voorzien.

Algemeen is men het erover eens, dat het loten van een element uit een populatie (dat is hier een welomschreven verzameling van elementen waaraan een p-variate meting kan worden verricht) adequaat wordt beschreven door een kansruimte met gelijke kansen. Dat is de reden dat in de leidraad de vraag werd gesteld of de waarnemingen te beschouwen zijn als aselecte trekking uit een populatie. Dit is vanzelfsprekend redelijk als de facto een loting is uitgevoerd. Ook wanneer dat niet het geval is, kan, bij het ontbreken van enige aanwijzing van met de kenmerken van het onderzoek verband houdende selectie, zo'n voorstelling niet onredelijk zijn. Het is van het grootste belang zich de vraag te stellen uit welke populatie men de waarnemingen als een steekproef wenst te beschouwen.

Beschouw eens de volgende voorbeelden:

a. Een onderzoeker wil de structuur van lichaamsmaten van aankomende middelbare scholieren onderzoeken. Om organisatorische redenen kan hij slechts één gehele eerste klas in een grote school onderzoeken. Hij hoort van de schoolleiding dat indeling in klassen indertijd op alfabetische volgorde heeft plaatsgevonden. Hij postuleert dat achternaam en lichaamsmaten geen verband met elkaar houden en hij beschouwt zijn proefpersonen als een aselecte trekking uit de populatie van bij die school aankomende scholieren. Dat lijkt niet onredelijk, ook al is hier niet werkelijk geloot. Zou hij zijn proefpersonen willen beschouwen als een aselecte trekking uit alle aankomende middelbare scholieren, dan zou hij zich bovendien het hoofd moeten breken over de vraag of een bepaalde school niet leerlingen uit bepaalde wijken of maatschappelijke lagen rekruteert, met het gevaar dat dat wel met lichaamsmaten verband houdt.

b. Een andere onderzoeker, aan dezelfde organisatorische beperkingen onderworpen, wil de struktuur van een aantal maten die verband houden met houding tegenover leren, school en discipline bij dezelfde scholieren onderzoeken. Hij onderkent evenwel, dat het klasseverband als zodanig een belangrijke invloed heeft op de houding van de individuele leerlingen tegenover deze zaken. Bijvoorbeeld, hij meent dat de houding van een "opinieleider" in de klas die van zijn klasgenoten sterk zal beïnvloeden. Voor hem is het beslist niet redelijk zijn onderzoekgroep als een aselecte steekproef te beschouwen: de selectie hangt samen met een der kenmerken waar het hem om gaat. Hij zal dus moeite moeten doen om genoemde organisatorische moeilijkheden te overwinnen, of moeten afzien van zijn voorgenomen analyse.

In de meeste gevallen waarin factoranalyse wordt gebruikt is het loten van elementen uit een populatie niet de enige bron van niet-gedetermineerdheid die in het statistisch model door een stochastische variabele wordt gerepresenteerd.

Vaak moet men aannemen, dat, wanneer eenmaal een zeker element uit een populatie is geselekteerd, herhaalde observaties aan de kenmerken van dit element geen constante uitkomsten zouden opleveren. Dikwijls worden deze bij herhaling te constateren verschillen geweten aan zgn. "storende invloeden". Men beschouwt de afwijking als geloot uit de verzameling van alle mogelijke afwijkingen op grond van storende invloeden.
Dit kan men beschrijven met een kansmodel:

$$\underline{y} = \underline{x} + \underline{e}$$

waarin alle variabelen p-variate stochastische variabelen zijn, en wel

$\underline{e}$ de vector van afwijkingen op grond van storende invloeden
$\underline{x}$ de vector van kenmerken van een element
$\underline{y}$ de vector van waarnemingen aan een element.

Het verdient daarbij de aandacht dat (waarbij, zonder beperking van algemeenheid wordt aangenomen, dat $E\underline{y}$, $E\underline{x}$ en $E\underline{e}$ nul zijn) de covariantie-

matrix $E\underline{yy}'$ van de te observeren scores, waar de hele factoranalyse om draait, als volgt is opgebouwd:

$$E\underline{yy}' = E\underline{xx}' + E\underline{xe}' + E\underline{ex}' + E\underline{ee}'.$$

Een postulaat dat $\underline{y}$ een factorstructuur heeft kan derhalve op diverse wijzen met de structuur van $\underline{x}$ en $\underline{e}$ samenhangen.

Meestentijds zal men in feite uit zijn op onderzoek van een eventuele factorstructuur van $\underline{x}$. Deze structuur weerspiegelt zich slechts ondubbelzinnig in die van $\underline{y}$ als de elementen van de vector $\underline{e}$ onderling èn van de elementen van de vector $\underline{x}$ stochastisch onafhankelijk zijn.

Terugvertaald naar de reële situatie betekent dat dat de storende invloeden geen verband mogen hebben met de grootte van de waar te nemen kenmerken en dat bovendien de inwerkingen van de storende invloeden op de verschillende kenmerken met elkaar geen verband houden. Al met al nogal zware eisen. Om deze reden kan het zeer nuttig zijn de gedachten even te laten verwijlen bij het mechanisme dat achter de data vermoed wordt.

Als toelichting de volgende voorbeelden, waarbij bovenstaande notatie is aangehouden:

a. Bij onderzoek naar de factorstructuur van een psychologische testbatterij worden aan een aselecte steekproef uit de onderzoekspopulatie de tests afgenomen. De onderzoeker in kwestie weet dat herhaalde afname van de tests aan een en dezelfde proefpersoon verschillende resultaten geeft. Hij wijt dit aan storende invloeden en meestentijds is daarmee de kous af. Bij enig navragen noemt hij als mogelijke storende invloed o.a. vermoeidheid van de proefpersoon. Nu is het zeer aannemelijk dat vermoeidheid de testscores op alle tests uit deze batterij drukt, zodat de elementen van de vector $\underline{e}$ niet langer ongecorreleerd zijn. Wanneer de variantie van de elementen van $\underline{e}$ fors is t.o.v. die van de elementen van $\underline{x}$, hetgeen in psychologische tests nogal eens het geval is, dan loopt deze psycholoog een goede kans een heel andere factorstructuur in zijn waarnemingen $\underline{y}$ te vinden dan er in $\underline{x}$ zit, terwijl hij eigenlijk naar de laatste op zoek is.

b. Een geval waar er sprake is van gecorreleerdheid van $\underline{x}$ en $\underline{e}$ is bijvoorbeeld het volgende:
scores komen tot stand door beoordeling van een reeks prestaties van een aselect gekozen groep proefpersonen. Als storende invloed wordt het samenstel der persoonlijke eigenaardigheden van de menselijke beoordelaar onderkend. Het is licht denkbaar dat men na beoordeling van de eerste kenmerken zich een bepaald beeld van de te beoordelen persoon heeft gevormd en daaraan, onbewust, zijn latere oordelen aanpast. Dit introduceert afhankelijkheid van de latere componenten van $\underline{e}$ met de eerste componenten van $\underline{x}$.

## 2.6. TWEE OF MEER POPULATIES

Wanneer men in feite beschikt over waarnemingen uit twee of meer populaties, waarvan men niet kan aannemen dat de verdeling van de kenmerken in elk der populaties gelijk is, dan zal men, wanneer men aan structuuronderzoek of datareductie denkt evenveel aparte analyses moeten uitvoeren als er populaties zijn.

Het wel gangbare gebruik de waarnemingen "op één hoop" te gooien en dan één analyse uit te voeren moet, wanneer geen nadere assumpties kunnen worden gerechtvaardigd, worden afgeraden.

Beschouw eens het volgende eenvoudige voorbeeld:
Zij $\underline{x}^{(1)}$ een multivariate stochastische variabele geassocieerd met het aselect trekken van een element uit populatie 1, en zij $\underline{x}^{(2)}$ de overeenkomstige variabele met populatie 2 geassocieerd.
Laat $\underline{x}^{(1)}$ en $\underline{x}^{(2)}$ alleen in verwachting verschillen en gelijke covariantiematrices hebben, die een m-factorontbinding toelaten. D.w.z.
$\underline{x}^{(1)} = \mu^{(1)} + \underline{y}^{(1)}$, $\underline{x}^{(2)} = \mu^{(2)} + \underline{y}^{(2)}$, met $\underline{y}^{(1)}$ en $\underline{y}^{(2)}$ gelijkverdeeld en $E\underline{y}^{(1)} = E\underline{y}^{(2)} = 0$ en $\text{cov}(\underline{x}^{(1)}) = \text{cov}\ (\underline{x}^{(2)})$.

Voor het voorbeeld nemen we $\text{cov}(\underline{x}^{(i)}) = I$, d.w.z. er is een nulfactorstructuur. Het "op één hoop" gooien van de variabelen komt neer op het beschouwen van een variabele $\underline{x} \overset{\text{def}}{=} \underline{x}^{(\underline{s})}$, waarbij $\underline{s}$ een stochastische variabele met waardebereik $\{1,2\}$ is, die met bepaalde kans de waarde 1 aanneemt. Voor het voorbeeld nemen we $P(\underline{s}=1) = \frac{1}{2}$. Voor de verwachting van $\underline{x}$ geldt dan:

$$\underline{Ex} = \tfrac{1}{2}\mu^{(1)} + \tfrac{1}{2}\mu^{(2)}$$

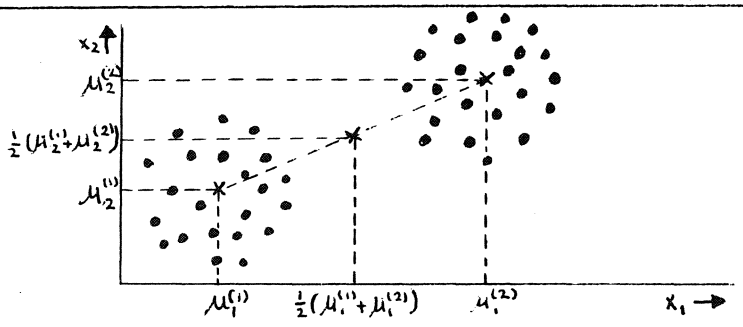



fig. 2.6.1 Twee populaties als één behandeld.

Factoranalyse tracht nu de afstanden van de diverse populatie-elementen t.o.v. het gemeenschappelijk middelpunt $\underline{Ex}$ te vatten in zo weinig mogelijk dimensies. Wanneer nu zoals in bovenstaande figuur, waar de verdeling van twee der componenten $x_1$ en $x_2$ is getekend, de afstand tussen de populatie-middelpunten $\mu^{(1)}$ en $\mu^{(2)}$ groot is t.o.v. de afstand tussen de scores van individuen behorend tot één populatie, dan introduceert de "op één hoop"-behandeling een factor:

Zij $\underline{s}^* \overset{\text{def}}{=} 2\underline{s} - 3$, dan geldt $E\underline{s}^* = 0$, var $(\underline{s}^*) = 1$ en de identiteiten

$$\underline{x}^{(1)} = \frac{\mu^{(1)}+\mu^{(2)}}{2} + \frac{\mu^{(2)}-\mu^{(1)}}{2} \cdot(-1) + \underline{y}^{(1)},$$

$$\underline{x}^{(2)} = \frac{\mu^{(1)}+\mu^{(2)}}{2} + \frac{\mu^{(2)}-\mu^{(1)}}{2} \cdot(+1) + \underline{y}^{(2)},$$

$$\underline{x} = \underline{x}^{(\underline{s})}.$$

Deze kunnen worden samengevat tot

$$\underline{x} = \frac{\mu^{(1)}+\mu^{(2)}}{2} + \frac{\mu^{(2)}-\mu^{(1)}}{2} \cdot\underline{s}^* + \underline{y}$$

(waarin $\underline{y}$ een stochastische variabele is met dezelfde verdeling als $\underline{y}^{(1)}$ en $\underline{y}^{(2)}$) en dit is juist een éénfactormodel.

Door het op een hoop gooien is in dit geval dus een factor geïntroduceerd en wel een factor die onderscheid maakt tussen het tot de ene of de andere populatie behoren.

Een en ander kan, wanneer $cov(\underline{x}^{(1)}) = cov(\underline{x}^{(2)})$, worden ondervangen door i.p.v. naar $\underline{x}$ naar $\underline{y}^{(s)}$ te kijken, d.w.z. naar de gereduceerde waarnemingen. Merk evenwel op, dat in de praktijk de benodigde veronderstelling $cov(\underline{x}^{(1)}) = (\underline{x}^{(2)})$ vaak niet vervuld zal zijn.

Wat er dan uit de bus zou komen is niet zonder meer te zeggen. Het samenvoegen zou dan ook in $\underline{x}^{(1)}$ en $\underline{x}^{(2)}$ aanwezige afzonderlijke en verschillende factorstructuren kunnen verdoezelen.

Voor een praktijkvoorbeeld zij verwezen naar 2.14. Deze behandelt Emmett's heranalyse van data van Slater en Benett over de aan- of afwezigheid van ruimtelijk inzicht bij 11-jarige schoolkinderen. Emmett zou bij de analyse onderscheid willen maken tussen jongens en meisjes. Dit bleek niet mogelijk. In dit geval zal het onderscheid tussen beide populaties vermoedelijk niet zo groot zijn als in het voorgaande voorbeeld. De factor "sekse" zal dus niet duidelijk als aparte factor gaan optreden, maar verweven raken met de andere factoren, met alle daaruit voortvloeiende moeilijkheden voor de interpretatie.

Al met al is factoranalyse een techniek die duidelijk bedoeld is voor het karakteriseren van één stochastische variabele, geassocieerd met één populatie.

## 2.7. SCHAALTYPE EN LINEAIRE SAMENHANG

Factoranalyse wordt gebruikt om de samenhang van variabelen te bestuderen. Hierbij moet wel worden opgemerkt dat factoranalyse slechts kan worden toegepast wanneer aan de variabelen redelijkerwijs een lineair model ten grondslag ligt. Dit is afhankelijk van het schaaltype van de variabelen.

Dit vindt zijn oorzaak in het feit dat factormodellen betrekking hebben op de covariantie- of product-moment correlatiestructuur van de variabelen in kwestie.

### *Schaaltype*

Men onderscheidt variabelen naar hun schaaltype.

a. *Nominale variabelen*

Een nominale variabele is een kenmerk dat de experimentele eenheden in groepen verdeelt, waarbij in een groep alle eenheden worden ondergebracht waarvan het kenmerk gelijk is. Wanneer twee experimentele eenheden in verschillende groepen terechtkomen, dan betekent dit slechts dat hun kenmerken verschillend zijn. Het zegt niets over de aard van het verschil. In het algemeen worden met de zo ontstane groepen kortheidshalve getallen geassocieerd om die groepen aan te duiden. Soms komen mensen in de verleiding met die getallen te gaan rekenen, maar dit is absurd. Deze getallen hebben geen andere betekenis dan een naamplaatje. De enige toegestane relaties zijn de identiteitsrelatie (=) en zijn ontkenning (≠).
Voorbeeld: godsdienst van een proefpersoon.

b. *Ordinale variabelen*

Bij ordinale variabelen gaat men een stap verder. Tussen de groepen van eenheden met gelijk kenmerk denkt men een zekere ordeing. Aan de groepen worden getallen toegekend die deze ordening symboliseren door groepen die lager in de ordening voorkomen, een kleiner getal toe te kennen. Op de zo ontstane getallen mogen behalve de relaties = en ≠ ook de groter- en kleinerrelaties ($<,\leq,>,\geq$) worden betrokken.

Rekenen met deze getallen kan zinvol zijn (zoals bij de toets van Wilcoxon), maar men dient de nodige voorzichtigheid te betrachten. Uit de getallen volgt de ordening, maar omgekeerd legt de ordening de getallen allesbehalve éénduidig vast. I.p.v. bijvoorbeeld de getallen 1,2,3,...,n kan men de kwadraten ervan nemen - dit levert dezelfde ordening op. Hieruit volgt dat het riskant is bij ordinale variabelen een lineair model te veronderstellen.
Voorbeeld: opleiding van proefpersonen (lager, middelbaar, hoger).

c. *Intervalvariabelen en ratiovariabelen*

Een intervalvariabele is een ordinale variabele waarbij het mogelijk wordt geacht de afstand tussen experimentele eenheden in een meeteenheid uit te drukken. Bij intervalvariabelen zijn behalve de bovengenoemde relaties ook de operaties optellen en aftrekken(+,-) zinvol.

Voorbeeld: geboortejaar van proefpersonen.

Rationiveau variabelen zijn intervalvariabelen waarbij een nulpunt is gedefinieerd, zodat het zinnig is te zeggen dat een waarde een aantal maal zo groot is als een andere waarde van deze variabele. Hier zijn de operaties vermenigvuldigen en delen (×,/) dus toegestaan.
Voorbeeld: leeftijd van proefpersonen.

Wanneer een intervalvariabele geen natuurlijk nulpunt heeft, kan het zinvol zijn het gemiddelde van deze variabele in een zekere populatie als nulpunt te beschouwen en zodoende een rationiveau variabele te creëren.

Daar factoranalyse uitgaat van de covariantie tussen de variabelen, of van de van covariantie afgeleide product-moment correlatiecoëfficiënt, en de covariantie een samenhangsmaat is die gebaseerd is op verschillen tussen de waar te nemen waarden en hun verwachting, moet het zinvol zijn bij die variabelen naar verschillen te kijken. Veelal zal dit het geval zijn bij interval- en rationiveau variabelen, soms ook bij ordinale variabelen.

Daar de covariantie niet gevoelig is voor locatieverschil, kan zonder beperking van algemeenheid voor alle variabelen worden aangenomen dat ze verwachting nul hebben: zij $\tilde{\underline{x}} = \underline{x} - E\underline{x}$, $\tilde{\underline{y}} = \underline{y} - E\underline{y}$ dan geldt $\mathrm{cov}(\tilde{\underline{x}},\tilde{\underline{y}}) = \mathrm{cov}(\underline{x},\underline{y})$.

Zonder beperking van algemeenheid betekent hier dat men de analyse veilig op de gereduceerde data $\tilde{\underline{x}},\tilde{\underline{y}}$ etc. kan uitvoeren. Bij interpretatie of gebruik van de resultaten moet men hiermee natuurlijk wel rekening houden. Vergelijk ook 2.6.

Factoranalyse is alleen toepasbaar bij lineaire samenhangen. Het klakkeloos toepassen van een lineair model kan in vele gevallen riskant zijn, want zelfs bij interval- en rationiveau variabelen kan de lineariteit een probleem vormen. Voor ordinale variabelen is een lineair model vaak riskant, maar het hangt van de feitelijke situatie af of het verondersteld mag worden. Als het goed past is het desalniettemin interpreteerbaar, al was het maar als benadering van onderliggende intervalvariabelen.

Samenvattend: voor nominale variabelen is een lineair model nooit bruikbaar, voor ordinale variabelen soms en voor interval- en rationiveau variabelen veelal.

*Lineaire samenhang*

De covariantie of de product-moment correlatiecoëfficiënt is een samenhangsmaat die geschikt is voor het karakteriseren van lineaire samenhang, waarmee bedoeld wordt dat elke puntenwolk van een steekproef van bivariate waarnemingen ruwweg rond een rechte lijn geconcentreerd is. Andere vormen van samenhang worden er niet of op onvoorspelbare wijze door weerspiegeld.

Wanneer tussen variabelen $x_1$ en $x_2$ het volgende kwadratische verband (fig. 2.7.1) bestaat, dan is de covariantie tussen $x_1$ en $x_2$ nul, en men krijgt dan, ten onrechte, de indruk dat er geen relatie bestaat. Is daarentegen de relatie, wederom kwadratisch, als in figuur 2.7.2, dan zal de product-moment correlatiecoëfficiënt nog vrij dicht bij 1 liggen. U bent dus nogal afhankelijk van het waardebereik van de variabelen in zulke gevallen.

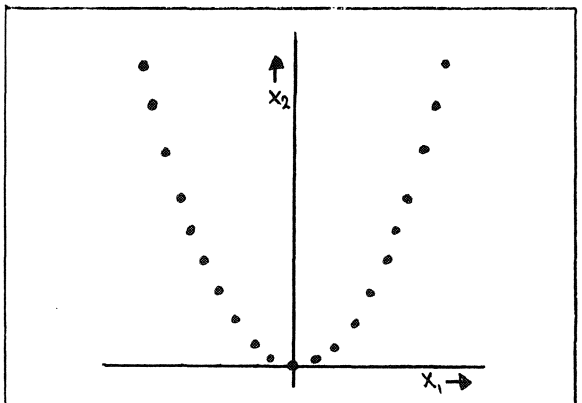



fig. 2.7.1

Een populatie van waarden $(x_1,x_2)$

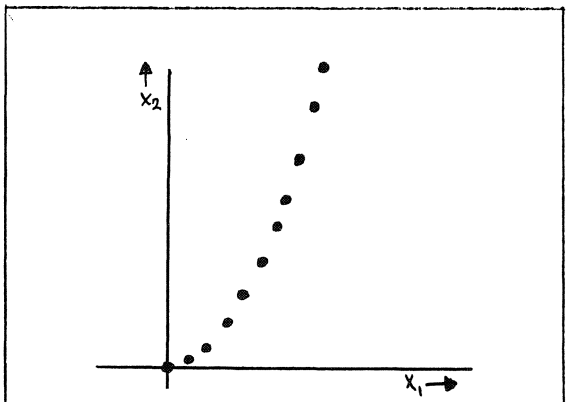



fig. 2.7.2

Een populatie van waarden $(x_1,x_2)$

Gebruik factoranalyse dus alleen als U redelijkerwijs aan kunt nemen dat er lineaire verbanden bestaan tussen de variabelen en ga dat in ieder geval na door het tekenen van puntenwolken. Eventueel kunt U door transformaties bepaalde verbanden lineariseren, maar bedenk wel dat een factorstructuur in getransformeerde data een factorstructuur in originele data veelal uitsluit, en omgekeerd.

Men dient goed te beseffen, dat de gevonden structuur dan geldt voor de getransformeerde variabelen.

## 2.8. OVER VERSCHIL TUSSEN PRINCIPALE COMPONENTENANALYSE EN FACTORANALYSE

Sommige beoefenaars van de multivariate analyse zijn van mening dat er geen wezenlijk verschil is tussen principale componentenanalyse (PCA) en factoranalyse (FA). Zij achten het twee manieren om hetzelfde te bereiken. Met "niet wezenlijk" wordt meestal bedoeld dat de numerieke resultaten van de twee methoden, toegepast op dezelfde covariantiematrix, vaak sterk op elkaar lijken.

Ons inziens is het wel van belang de beide benaderingen te onderscheiden. In de eerste plaats beogen ze iets anders. In de tweede plaats kunnen er ook grote verschillen tussen numerieke resultaten voorkomen. We zullen dit nagaan door eerst globaal PCA te beschrijven en vervolgens een eenvoudig voorbeeld van dergelijke verschillen te geven.

*PCA*

We geven hier een korte karakterisering van PCA. Voor een meer gedetailleerde uiteenzetting over het uitvoeren van PCA zie bijv. MORRISON (1976).

PCA is een techniek die tracht de positie van alle experimentele eenheden in de totale puntenwolk, die bepaald wordt door de waarde van de p variabelen voor iedere eenheid, op zuiniger wijze te karakteriseren. Dit gebeurt door een aantal, zeg m, onderling ongecorreleerde lineaire combinaties $\underline{\pi}_1,\ldots,\underline{\pi}_m$ van de originele variabelen $\underline{x}_1,\ldots,\underline{x}_p$ te construeren met de eigenschap dat $\underline{x}_1,\ldots,\underline{x}_p$ zo goed mogelijk op basis van $\underline{\pi}_1,\ldots,\underline{\pi}_m$ kunnen worden gereconstrueerd. D.w.z. dat $\underline{\pi}_1,\ldots,\underline{\pi}_m$ de beste lineaire voorspellers van $\underline{x}_1,\ldots,\underline{x}_p$ moeten zijn, ofwel dat, als $\sigma_i^2$ de residuele variantie bij lineaire voorspelling van $\underline{x}_i$ op grond van $\underline{\pi}_1,\ldots,\underline{\pi}_m$ is, dat dan $\sum_{i=1}^p \sigma_i^2$ minimaal moet zijn onder alle keuzen van m lineaire combinaties $\underline{\pi}_1,\ldots,\underline{\pi}_m$ van $\underline{x}_1,\ldots,\underline{x}_p$.
Men is gewend om niet meer naar $\sum_{i=1}^p \sigma_i^2$ te kijken, maar naar zijn tegenhanger $\sum_{i=1}^p \mathrm{var}(\underline{x}_i) - \sum_{i=1}^p \sigma_i^2$. Deze grootheid noemt men de verklaarde variantie; veelal wordt de verklaarde variantie ook uitgedrukt als percentage van $\sum_{i=1}^p \mathrm{var}(\underline{x}_i)$.

Wanneer we het bovenstaande uitdrukken in termen van de geobserveerde waarden $x_{ki}$, de waarde van de i-de variabele voor de k-de eenheid, dan betekent dat dat gezocht wordt naar constanten $\mu_i$, $\lambda_{ij}$, $\pi_{kj}$ ($i = 1,\ldots,p$; $j = 1,\ldots,m$; $k = 1,\ldots,n$ met n het aantal experimentele eenheden) zodanig dat

$$\sum_{i=1}^{p} \sum_{k=1}^{n} (x_{ki} - \mu_i - \sum_{j=1}^{m} \lambda_{ij}\pi_{kj})^2 \qquad \text{minimaal is.}$$

De variabelen $\underline{\pi}_1,\ldots,\underline{\pi}_m$ (resp. de vectoren $(\underline{\pi}_{11},\ldots,\underline{\pi}_{n1}),\ldots,(\underline{\pi}_{1m},\ldots,\underline{\pi}_{nm})$) heten principale componenten. De constanten $\mu_i$ blijken de gemiddelden van $x_{1i},\ldots,x_{ni}$ te zijn.

Hoe groot men m, het aantal componenten, kiest is arbitrair. Men laat dit ervan afhangen of men de verklaarde variantie groot genoeg acht. Als $\underline{x}_1,\ldots,\underline{x}_p$ niet lineair afhankelijk zijn heeft men p componenten nodig om de verklaarde variantie tot 100% te laten stijgen.

Over de berekening van principale componenten vermelden we slechts het volgende. Laat $\Sigma$ de covariantiematrix van $\underline{x} = (\underline{x}_1,\ldots,\underline{x}_p)'$ zijn, met eigenwaarden $\lambda_1,\ldots,\lambda_p$ in aflopende grootte en bijbehorende genormaliseerde eigenvectoren $\omega_1,\ldots,\omega_p$, dan zijn de principale componenten te bepalen als

$$\omega_1' \underline{x},\ \omega_2' \underline{x},\ldots,\ \omega_p' \underline{x}$$

met als door de eerste m componenten verklaarde variantie de waarde $\sum_{i=1}^{m} \lambda_i$.

Merk op dat deze techniek om de plaats van de experimentele eenheden binnen de puntenwolk te karakteriseren door het gereduceerde aantal van m grootheden i.p.v. de originele p, voor elk willekeurig waarnemingsmateriaal uitvoerbaar is, voor elke waarde van $m \leq p$. Er ligt geen restrictief model aan ten grondslag. Slechts het aantal componenten nodig om een zekere -arbitrair gekozen- hoeveelheid verklaarde variantie te bereiken is wisselend.

*Is PCA hetzelfde als FA?*

In het algemeen verschilt PCA al hierin van FA dat het een uitvoerbare techniek is, onafhankelijk van het feit of er een factormodel aan de data ten grondslag ligt (of niet). Bij factoranalyse bestaat althans het

-soms ook gerealiseerde- streven d.m.v. toetsen de geldigheid van het model te onderzoeken. Dit is een belangrijk verschil. Maar stel dat een stochastische variabele $\underline{x}$ een m-factormodel heeft d.w.z. er geldt $\underline{x} = \mu + \Lambda_F\underline{\phi} + \underline{\varepsilon}$ met de gebruikelijke veronderstellingen. Zij verder $\underline{\pi} \overset{def}{=} (\pi_1,\ldots,\pi_m)'$ de vector van de eerste m principale componenten met coëfficiëntenmatrix $\Lambda_p$, d.w.z. $\underline{x} = \mu + \Lambda_P\underline{\pi} + \underline{r}$, voor zekere residuenvector $\underline{r}$. Geldt dan misschien dat $\Lambda_F = \Lambda_P$ en $\underline{\phi} = \underline{\pi}$ en $\underline{\varepsilon} = \underline{r}$?

Het zou tenslotte zo kunnen zijn dat twee technieken met verschillende doeleinden dezelfde resultaten geven.

Dit is echter niet zo. $\underline{\pi}$ en $\underline{\phi}$ zijn niet gelijk, omdat $\underline{\pi}$ eenduidig bepaald is door de geobserveerde covariantiematrix $\Sigma$, maar $\underline{\phi}$ niet (zie 2.12). $\underline{\varepsilon}$ en $\underline{r}$ zijn niet gelijk, daar, wanneer $\Sigma$ niet singulier is, $\text{cov}(\underline{r})$ geen diagonaalmatrix is en $\text{cov}(\underline{\varepsilon})$ wel. Daar $\Lambda_F\Lambda_F' = \Sigma - \text{cov}(\underline{\varepsilon})$ en $\Lambda_P\Lambda_P' = \Sigma - \text{cov}(\underline{r})$, kunnen ook $\Lambda_F$ en $\Lambda_P$ niet gelijk zijn.

*Is PCA bijna hetzelfde als FA?*

De proponenten van de visie dat PCA en FA niet wezenlijk verschillen zullen door het bovenstaande niet geschokt zijn omdat ze niet volhouden dat PCA en FA precies hetzelfde zijn. Zij vinden dat de verschillen in numerieke uitkomsten niet wezenlijk zijn, d.w.z. dat $\Lambda_F$ en $\Lambda_P$ erg veel op elkaar lijken. Of dit juist is hangt natuurlijk erg af van wat men onder niet wezenlijk verstaat. Wij geven hier een voorbeeld waarin wij de verschillen tussen beide matrices wel wezenlijk vinden.

Daartoe beschouwen wij de PCA van een GFA-model omdat GFA qua oplossingsmethode het nauwst aansluit bij PCA. Zij gegeven de covariantiematrix $\Sigma$ van een stochastische variabele $(\underline{x}_1,\underline{x}_2,\underline{x}_3)'$

$$\Sigma = \begin{bmatrix} 1 & r & r \\ r & 1 & r \\ r & r & 1 \end{bmatrix}, \quad 0 < r < 1$$

Deze voldoet aan een één-factor GFA-model $\Sigma = \Lambda_F\Lambda_F' + \Psi$.

$$\Lambda_F = (\sqrt{r}, \sqrt{r}, \sqrt{r})' \text{ en } \Psi = \begin{bmatrix} 1-r & 0 & 0 \\ 0 & 1-r & 0 \\ 0 & 0 & 1-r \end{bmatrix}.$$

De eerste principale component van deze matrix geeft als resultaat een ontbinding van $\Sigma$ als

$$\Sigma = \Lambda_P \Lambda_P' + R$$

met

$$\Lambda_P = \sqrt{1 + 2r}(1/\sqrt{3},\ 1/\sqrt{3},\ 1/\sqrt{3})'$$

d.w.z. dat de eerste principale component evenredig is met $\underline{x}_1 + \underline{x}_2 + \underline{x}_3$, en dat slechts de evenredigheidsfactor afhankelijk is van r.
De residuele matrix R wordt gegeven door

$$R = \frac{1 - r}{3}\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}$$ en is dus niet diagonaal.

Het percentage verklaarde variantie is gelijk aan $\frac{1 + 2r}{3} \times 100\%$, hetgeen natuurlijk groter is dan de overeenkomstige grootheid in de GFA-ontbinding. Deze is de som van de varianties der gemeenschappelijke delen als percentage van de som van de varianties der $\underline{x}_i$ en is gelijk aan $r \times 100\%$.

De correlatiecoëfficiënt tussen de echte factor en de eerste principale component is gelijk aan $\sqrt{\frac{3r}{1 + 2r}}$, hetgeen er al op wijst dat factor en component flink kunnen verschillen, met name als r niet groot is.
Laten we dit bekijken voor het geval $r = \frac{1}{4}$. Dan geldt

$$\Sigma = \begin{bmatrix} 1 & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & 1 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & 1 \end{bmatrix}, \qquad \Lambda_F = (\tfrac{1}{2},\ \tfrac{1}{2},\ \tfrac{1}{2})'$$

$$R = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{1}{2} \end{bmatrix}, \qquad \Lambda_P = (\sqrt{\tfrac{1}{2}},\ \sqrt{\tfrac{1}{2}},\ \sqrt{\tfrac{1}{2}})'$$

D.w.z. dat de absolute waarde van de niet-diagonaal elementen van $\Sigma$ in R niet is veranderd. Als covariantieverklaring is deze PCA-oplossing zeker geen succes. Ook verschillen $\Lambda_F$ en $\Lambda_p$ flink. De verklaarde variantie bedraagt 50% bij de PCA-oplossing en 25% bij de GFA-oplossing.

Tenslotte verschillen factor en component nog wel zoveel dat hun correlatiecoë-ficiënt gelijk is aan $\sqrt{\frac{1}{2}}$ ofwel ongeveer 0,7.

Tezamen vinden wij dat de beide oplossingen wezenlijk verschillen.

Wanneer lijken PCA- en FA-oplossingen nu wel op elkaar?. Dat is ons niet zo duidelijk; in ieder geval wel als een factormodel geldt waarin alle communaliteiten zeer hoog zijn.

## 2.9. SCHAALAFHANKELIJKHEID

In veel gevallen waarin factoranalyse wordt toegepast hebben de meeteenheden waarin de variabelen worden uitgedrukt geen intrinsieke betekenis. Dan is het noodzakelijk dat de factormodellen bestand zijn tegen lineaire schaaltransformaties, d.w.z. als de lineaire samenhang van $\underline{x}_1,\ldots,\underline{x}_p$ wordt verklaard door hun afhankelijkheid van factoren $\underline{\phi}_1,\ldots,\underline{\phi}_m$ met nadere eisen aan de uniciteiten, dan moet de lineaire samenhang van $a_1\underline{x}_1,\ldots,a_p\underline{x}_p$ kunnen worden verklaard op grond van *dezelfde* $\underline{\phi}_1,\ldots,\underline{\phi}_m$ met dezelfde eisen aan de uniciteiten, voor elk stel positieve getallen $a_1,\ldots,a_p$.

Wij onderzoeken dat hier voor de verschillende modellen. Laat $\underline{x}$ een covariantiematrix $\Sigma$ en een factormodel $\Sigma = \Lambda\Lambda' + \Psi$ hebben.
Zij $A = \begin{bmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_p \end{bmatrix}$ een matrix van schaaltransformaties met positieve diagonaalelementen.

De getransformeerde variabelen $A\underline{x}$ hebben covariantiematrix $A\Sigma A'$ die de ontbinding

$$A\Sigma A' = (A\Lambda)(A\Lambda)' + \Psi\cdot A^2$$

toelaat. De verklaring van de covarianties met dezelfde factoren (maar andere factorladingen) is dus nooit een probleem. De crux ligt in de uniciteitenmatrix $\Psi\cdot A^2$.

Onder AFA wordt slechts de eis gesteld dat de uniciteitenmatrix een positieve diagonaalmatrix is; als $\Psi$ daaraan voldoet dan voldoet $\Psi A^2$ daar ook aan.

Ook de eis van JFA aan de uniciteitenmatrix blijft behouden:

als $\Psi = \theta(\text{diag}(\Sigma^{-1}))^{-1}$

dan geldt

$$\Psi A^2 = \theta A(\text{diag}(\Sigma^{-1}))^{-1}A = \theta(A^{-1}\text{diag}(\Sigma^{-1})A^{-1})^{-1} =$$

$$= \theta(\text{diag}(A^{-1}\Sigma^{-1}A^{-1}))^{-1} =$$

$$= \theta(\text{diag}((A\Sigma A)^{-1}))^{-1} \quad \text{q.e.d.}$$

Heel anders ligt de zaak in het geval GFA:

als $\Psi = \sigma^2 I$

dan geldt

$$\Psi A^2 = \sigma^2 A^2 \neq \tau^2 I$$

tenzij A zelf van de vorm aI is.

Het gelijke residuele variantiesmodel is dus niet bestand tegen schaaltransformaties die niet voor elke variabele gelijk zijn. Dit impliceert dat het niet zinvol is het gelijke residuele variantiesmodel toe te passen wanneer de variabelen niet zijn uitgedrukt in een meeteenhedenstelsel met een voor dat probleem intrinsieke betekenis (behoudens gelijke schaaltransformaties voor alle variabelen tegelijkertijd).

*Schaalafhankelijkheid van principale componenten*

Aan de hand van een eenvoudig voorbeeld zullen wij de gevoeligheid voor schaaltransformaties van principale componentenanalyse laten zien. (Een uitgebreider behandeling wordt gegeven in 3.3.) Dit voorbeeld behan-

delt het geval van een bivariate stochastische variabele $\underline{x} \overset{def}{=} (\underline{x}_1, \underline{x}_2)'$ met covariantiematrix

$$\Sigma = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}, \text{ met } r > 0.$$

Beschouw nu de bivariate variabelen $\underline{x}(a)$ die ontstaan door schaaltransformaties toe te passen op alléén de eerste variabele $\underline{x}_1$ met een positieve constante a, d.w.z. we beschouwen $\underline{y}(a) = (\underline{y}_1(a), \underline{y}_2(a))' \overset{def}{=} (a\underline{x}_1, \underline{x}_2)'$, met covariantiematrix

$$\Sigma_a \overset{def}{=} \text{cov}(\underline{y}(a)) = \begin{bmatrix} a^2 & ar \\ ar & 1 \end{bmatrix}, \quad a \in (0,\infty).$$

De eerste principale component $\underline{z}_{(a)}$ is de lineaire combinatie

$$\underline{z}(a) = \sqrt{1 - \lambda^2} \cdot \underline{y}_1(a) + \lambda \cdot \underline{y}_2(a)$$

die onder alle mogelijke keuzen van $\lambda$ uit $[-1, 1]$ de grootste variantie heeft.

$\lambda$ is een functie van a en we zullen zien dat elke waarde uit (0,1) door een geschikte keuze van a gerealiseerd kan worden. Dat betekent dat men door keuze van een geschikte schaaltransformatie elke willekeurige lineaire combinatie van de getransformeerde variabelen als principale component kan verkrijgen, ongeacht hoe het waarnemingsmateriaal eruit ziet. Immers, de grootste eigenwaarde van $\Sigma_a$ is

$$\delta_1(a) = \tfrac{1}{2}(a^2 + 1 + D(a))$$

waarin

$$D(a) \overset{def}{=} \{(a^2 - 1)^2 + 4a^2r^2\}^{\frac{1}{2}}.$$

De bijbehorende genormaliseerde eigenvector is dan

$$\omega_1(a) = \{(a^2 - 1 + D(a))^2 \, (2ar)^{-2} + 1\}^{-\frac{1}{2}} \cdot \begin{bmatrix} (a^2 - 1 + D(a))/2ar \\ 1 \end{bmatrix}$$

en

$$\underline{z}(a) = \omega_1(a)' \cdot \underline{y}(a), \qquad \text{d.w.z.}$$

de coëfficiënt $\lambda$ van $\underline{y}_2$ in de lineaire combinatie die $\underline{z}(a)$ is, is gelijk aan

$$\lambda = \{(a^2 - 1 + D(a))^2 \cdot (2ar)^{-2} + 1\}^{-\frac{1}{2}}.$$

Enige algebra laat zien dat $\lim_{a \downarrow 0} \lambda = 1$ en $\lim_{a \to \infty} \lambda = 0$ en daar $\lambda$ een continue functie van a op $(0,\infty)$ is, is daarmee aangetoond dat, door geschikte keuze van a, $\lambda$ elke waarde uit (0,1) kan aannemen.

Ter nadere illustratie vermelden we nog dat de fractie door $\underline{z}(a)$ verklaarde variantie f(a) gelijk is aan

$$f(a) = \tfrac{1}{2}\{1 + D(a) \cdot (a^2 + 1)^{-1}\}.$$

Hiervoor geldt $\lim_{a \to \infty} f(a) = \lim_{a \downarrow 0} f(a) = 1$, d.w.z. dat passende schaaltransformatie de fractie verklaarde variantie willekeurig hoog kan opvoeren. De correlatiecoëfficiënt tussen $\underline{z}(a)$ en $\underline{y}_2(a)$ wordt gegeven door

$$\rho(\underline{z}(a), \underline{y}_2(a)) = \{\tfrac{1}{2}(a^2 + 1 + D(a))\}^{\frac{1}{2}} \cdot \{(a^2 - 1 + D(a))^2/4a^2r^2 + 1\}^{-\frac{1}{2}}$$

en deze nadert tot 1 als $a \to 0$ en tot r als $a \to \infty$, zoals overeenkomt met het feit dat $\underline{z}(a)$ dan geheel door $\underline{y}_2(a)$ respectievelijk $\underline{y}_1(a)$ wordt bepaald.

Samenvattend betekent dit dat principale componentenanalyse, zoals te verwachten is bij een variantiegerichte techniek, zeer schaalgevoelig is. Voor een datareductietechniek hoeft dat geen bezwaar te zijn, maar het is goed zich te realiseren hoezeer de resultaten van een analyse van de schaalkeuze afhankelijk zijn. Immers de gekozen schaal behoeft niet altijd de voor datareductie meest succesvolle schaal te zijn.

*Standaardiseren*

Vaak wordt standaardiseren van variabelen aanbevolen om de schaalproblemen bij daarvoor gevoelige technieken op te lossen. (Standaardiseren heeft tot gevolg dat niet de covariantiematrix maar de correlatiematrix aan een analyse wordt onderworpen.) Eigenlijk is dit niet zozeer een oplossen van het probleem alswel het op conventionele manier één schaal uitkiezen. Bij principale componentenanalyse hoeft dit niet echter altijd de voor datareductie meest succesvolle schaal te zijn.

Bij de andere schaalgevoelige methode, GFA, heeft deze conventie bovendien een bijzondere betekenis, namelijk dat alle variabelen in gelijke mate door de factoren bepaald zouden worden (zie 1.4.).

## 2.10. HET AANTAL FACTOREN

Wanneer niet tevoren is vastgesteld hoeveel factoren het materiaal voortbrengen, dus in geval van de exploratieve variant, doet zich de vraag voor: hoeveel factoren zijn nodig om de data voort te brengen?

Nu is het zaak goed onderscheid te maken tussen twee vragen:

a. hoeveel factoren zijn nodig om de samenhang te verklaren?
b. hoeveel (en welke) factoren hiervan hebben praktisch nut?

Ten onrechte wordt vaak om het aantal factoren te bepalen alleen de tweede vraag beantwoord.

Ons inziens is de juiste weg eerst de eerste vraag op grond van statistische criteria te beantwoorden en vervolgens de tweede vraag te bekijken. Laat men het eerste na, dan komt men slechts tot vage uitspraken, daar men dan geen criterium heeft om de afwijkingen van de realiteit voor het gevonden model te beoordelen. De uitspraak "m factoren verklaren de samenhang grotendeels" is te vaag om empirische relevantie te hebben.

*De sequentiële methode*

De gangbare methode om het aantal factoren te schatten is als volgt: Men postuleert dat $\underline{x}$ multivariaat normaal verdeeld is (zie 2.11 of 1.8).

Als $H_m$ de hypothese is dat een m-factormodel geldt dan toetst men achtereenvolgens $H_0$ (0-factormodel), $H_1$ (1-factormodel), $H_2$,... net zolang tot men een waarde $m_0$ heeft gevonden zodanig dat $H_{m_0-1}$ wel verworpen wordt, maar $H_{m_0}$ niet.

De waarde $m_0$ is de schatting voor het aantal factoren. (Wanneer men geen $H_m$ kan verwerpen voordat m de restrictiviteitsgrens voor het betreffende model overschrijdt (zie 2.4), concludeert men dat geen factormodel geldt.)

De statistische eigenschappen van deze schattingsprocedure zijn zeer ingewikkeld omdat de toetsen voor $H_0$, $H_1$,..., $H_{m_0}$ niet onafhankelijk zijn. Wanneer men elk van deze aparte toetsen uitvoert met onbetrouwbaarheid $\alpha$, dan is de kans dat men een te grote waarde voor m vindt (veel) kleiner dan $\alpha$. Over de kans dat men een te kleine waarde vindt valt weinig te zeggen. Deze is in ieder geval sterk afhankelijk van het aantal waarnemingen.

Het komt zelfs vrij geregeld voor dat men $H_m$ voor zekere m niet verwerpt maar $H_{m+1}$ wel, zo men deze laatste hypothese nog zou toetsen. Wat we hier mee aanmoeten is niet duidelijk; in ieder geval maakt het de volgende raad aanbevelenswaardig.

Om echt vertrouwen in de zo geschatte waarde van m te krijgen zal men, met nieuw materiaal, deze waarde in de "m-factorvariant" dienen te toetsen. Wanneer men veel waarnemingen heeft kan men eventueel m schatten met de ene helft van de data en de gevonden waarde toetsen op de andere helft.

Voor de exacte beschrijving van de afzonderlijke toetsen $H_m$, zie 3.1.

In plaats van deze toetsingsprocedure zijn een aantal andere criteria in omloop, waarvan we er hier twee bespreken.

*Knikcriterium*

Dikwijls wordt aangeraden het aantal factoren te bepalen door een grafiekje van "met de factoren geassocieerde eigenwaarden" te tekenen. Nu is deze term beslist niet eenduidig en het is soms niet eenvoudig er achter te komen wat een gegeven factoranalyseprogramma precies vermeldt onder dit hoofdje.

Er zijn tenminste vier mogelijkheden:

a. eigenwaarden van $\Sigma$
b. eigenwaarden van $\Sigma^* \stackrel{def}{=} (\text{diag } \Sigma^{-1})^{-\frac{1}{2}} \Sigma (\text{diag } \Sigma^{-1})^{-\frac{1}{2}}$
c. eigenwaarden van $\Lambda\Lambda'$; deze zijn gelijk aan $\sum_{i=1}^{p} \lambda_{ij}^2$, $j = 1,\ldots,m$; en worden ook wel "door de factoren verklaarde variantie" genoemd
d. eigenwaarden van $\Psi^{-\frac{1}{2}}\Sigma\Psi^{-\frac{1}{2}}$

c. en d. zijn eerst beschikbaar nadat de analyse is uitgevoerd. c. behandelen we bij het volgende criterium. d. vormt in feite de basis voor de formele modeltoets en voor het Guttman criterium (zie 3.1 en 2.12). a. en b. bespreken we hier.
De hier bedoelde procedure vraagt een grafiekje als fig. 2.10.1 te tekenen van de onder a. vermelde eigenwaarden, uitgezet tegen hun rangnummer van grootste naar kleinste. Het aantal factoren wordt m.b.v. het plaatje geschat als het rangnummer waarna alle eigenwaarden ongeveer even groot zijn.

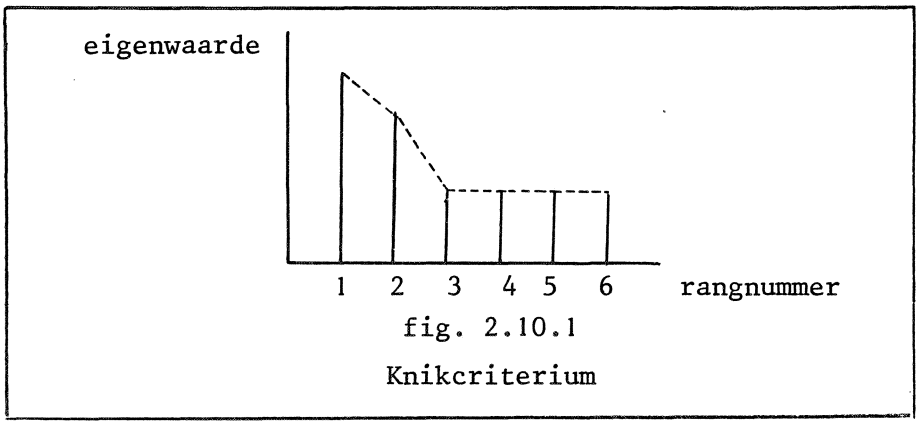



fig. 2.10.1
Knikcriterium

In fig. 2.10.1 zou dit aantal dus op 2 geschat zijn. Men spreekt van het knikcriterium omdat bij het volgende rangnummer een lijn over de toppen van de eigenwaardenbalkjes in het plaatje een knik vertoont.

Voor GFA is deze regel een informele variant van de eerder beschreven toetsingsprocedure. Voor AFA is de procedure echter niet correct en hij leidt vaak tot een onjuist aantal factoren of tot verwarring omdat geen eigenwaarden ongeveer gelijk schijnen.

Voor JFA is een variant van de procedure te hanteren. Men dient dan echter de eigenwaarden (b) van $\Sigma^*$ te gebruiken.

Zowel voor AFA, JFA als GFA geven wij de voorkeur aan de toets, indien toepasbaar. De informele methode kan echter zijn nut hebben voor een eerste inspectie van het materiaal.

*Eigenwaarden-groter-dan-1-criterium*

In de literatuur wordt dikwijls aangeraden het aantal factoren te bepalen als het aantal eigenwaarden van $\Sigma$ dat groter is dan 1.

Dit is eigenlijk geen criterium om een passend factormodel te vinden, maar een criterium om te bepalen hoeveel factoren practisch nut hebben. Het ontleent zijn betekenis aan PCA. Het criterium is alleen van toepassing wanneer $\Sigma$ een correlatiematrix is. Het berust op het feit, dat de j-de eigenwaarde van $\Sigma$ soms een aardige benadering vormt voor de "hoeveelheid door de j-de factor verklaarde variantie" (d.w.z. dat de eigenwaarden (a) soms niet veel verschillen van de eigenwaarden (c)).

Deze laatste grootheid is veelzeggend wanneer "praktisch nuttig" als "flinke variantie bezittend" vertaald wordt, bijvoorbeeld in gevallen waar de factoren bedoeld zijn om experimentele eenheden te differentiëren. Het criterium is i.h.a. passend voor datareductiegevallen, bijvoorbeeld in principale componentenanalyse.

Nu is de benadering van de eigenwaarden (c) door de eigenwaarden (a) niet zelden nogal pover. Men wordt aangeraden dit na de analyse te vergelijken en wanneer duidelijke verschillen optreden de analyse te herhalen met een ander aantal factoren.

Een variant van het groter-dan-1-criterium is het groter-dan-5%-criterium, d.w.z. zoveel factoren van belang te achten dat de laatste nog tenminste 5% van de totale variantie verklaart. Ook dit wordt dan met de eigenwaarde van $\Sigma$ benaderd: neem zoveel factoren als er eigenwaarden groter dan $\frac{1}{20}$ p zijn. Ook dit criterium is alleen van toepassing op correlatiematrices in datareductieproblemen. Dit criterium is zeer arbitrair, zoals uit het volgende blijkt. De som van de p eigenwaarden is gelijk aan p. Is nu het aantal variabelen groter dan 20, dan kan het gebeuren dat er geen enkele eigenwaarde groter dan $\frac{1}{20}$ p is. Dit hangt sterk van het aantal variabelen af.

Nogmaals wijzen wij erop, dat deze twee criteria (groter-dan-1 en groter-dan-5%) geen betrekking hebben op de vraag hoeveel factoren moeten worden gepostuleerd om een factormodel geldig te maken, maar op de vraag welke factoren practisch nut hebben.

*Praktisch nut*

Wanneer is vastgesteld of een factormodel geldt en met welk aantal factoren, komt de tweede vraag aan de orde: welke factoren hebben practisch nut. Het antwoord hierop hangt vanzelfsprekend sterk af van wat men met die factoren wil. Het vorige gedeelte gaf een voorbeeld in een datareductie-geval. In echte factoranalysegevallen geldt dat men alleen iets met factoren kan beginnen die aan het Guttman criterium voldoen (zie 2.12). Het aantal factoren dat daaraan voldoet is meestal aanzienlijk kleiner dan $m_0$.

## 2.11. TOETSEN, NAUWKEURIGHEID VAN SCHATTINGEN EN DE ROL VAN MULTIVARIATE NORMALITEIT

*Toetsen*

Wanneer wij de bewering dat m factoren de samenhang tussen de variabelen verklaren niet onderwerpen aan een statistische toets, dan bezit deze bewering geen enkele bewijskracht. Bij herhaling van het experiment zouden we mogelijk een volstrekt ander resultaat verkrijgen; zonder toetsing valt daarover niets te voorspellen. Het is onverantwoord te stellen dat de covariantie- of correlatiematrix goed verklaard wordt omdat we daarvoor geen maatstaf hebben. Een toets verschaft die maatstaf.

*Nauwkeurigheid van schattingen*

Wanneer we eenmaal door toetsing een zeker vertrouwen hebben in het gespecificeerde aantal factoren, is het vervolgens van belang de nauwkeurigheid van de schattingen van factorladingen en communaliteiten na te gaan. Ook hier loopt men weer het risico, zo men dit nalaat, dat herhaling van het experiment onverwacht grote verschillen in deze schattingen zou opleveren en daarmee een volstrekt andere interpretatie van de resultaten. Het feit, dat dit berekenen van standaardafwijkingen van schatters tot voor kort zelden plaatsvond, is ons inziens debet aan veel misplaatst vertrouwen in factoranalytische resultaten.

Wij bevelen dan ook ten sterkste aan, dat men aandacht besteedt aan de nauwkeurigheid van schattingen.

*Multivariate normaliteit*

Om met de thans bruikbare technieken toetsen uit te kunnen voeren en de nauwkeurigheid van schattingen te kunnen bepalen, moet aan een aantal voorwaarden worden voldaan, die in de LEIDRAAD de revue passeren. Een van de lastigste is misschien wel de vraag of de waarnemingen een multivariaat normale verdeling hebben. Hoe kan men nagaan of dit het geval is en wat valt te zeggen over de kwestie of een eventuele afwijking van deze eis te ernstig is om de resultaten van de normale theorie nog bij benadering geldig te achten?

*Nagaan van multivariate normaliteit*

Ondanks het belang van de vraag naar multivariate normaliteit is er door statistici nog bitter weinig concreets voorgesteld om deze aanname te toetsen.

Het wordt natuurlijk ten sterkste aanbevolen om in ieder geval per variabele univariate normaliteit te bekijken: men kan bijvoorbeeld voor iedere variabele een histogram van de waarnemingen maken, dat dan de bekende klokvorm zou moeten vertonen en men kan een normaliteitstoets uitvoeren (univariate normaliteitstoetsen en grafische procedures voor het nagaan van normaliteit, bijv. m.b.v. waarschijnlijkheidspapier, vindt men o.a. bij HEGAZY & GREEN (1975), HEMELRIJK & KRIENS (1967), ANSCOMBE & TUKEY (1963)).

Wanneer duidelijk aanwijzingen bestaan dat men met niet-normale verdelingen te doen heeft, is enig voorbehoud m.b.t. de berekende standaardafwijkingen van parameterschatters op zijn plaats, terwijl voor de toetsen een aanzienlijk groter voorbehoud in acht dient te worden genomen.

Univariate normaliteit van elke variabele apart garandeert nog geenszins multivariate normaliteit van alle variabelen tezamen.

Zo verkeren wij in de netelige positie een aanname te moeten doen die, voorzover ons bekend, niet op bevredigende manier te toetsen is. Het probleem staat gelukkig wel in de belangstelling (zie bijvoorbeeld ANDREWS, GNANADESIKAN & WARNER (1973)).

*Afwijkingen van multivariate normaliteit*

Hoe erg zijn afwijkingen van multivariate normaliteit? Het antwoord luidt eigenlijk: dat weten we niet. Alles wat we kunnen bieden zijn enige overwegingen over wat er mis kan gaan. In ieder geval menen we dat men, toch de op normaliteit gebaseerde standaardafwijkingen moet bepalen, ook wanneer de normaliteitseis niet vervuld lijkt te zijn. Beter een onnauwkeurige indruk van de onnauwkeurigheid van schatters dan helemaal geen indruk.

Wat kan er aan de hand zijn?

In de eerste plaats is het mogelijk dat wel een factormodel geldt, maar met niet-normaal verldeelde data. Factoranalyse maakt gebruik van de geschatte covariantiematrix van de variabelen. Wanneer deze schatters onnauwkeuriger zijn dan in geval van normaliteit dan zullen ook de factoranalyseschatters onnauwkeuriger zijn. Dit kan onder meer voorkomen als de betrokken variabelen extreme waarden met grotere kans aannemen dan onder normaliteit het geval zou zijn (zogenaamde verdelingen met dikke staarten). Dit kan er toe leiden dat we ofwel te veel vertrouwen hebben in een gevonden structuur ofwel ten onrechte menen dat de gevonden structuur te sterk afwijkt van wat we interpreteerbaar achten.

Een dergelijke situatie kan ook leiden tot onjuiste toetsingsuitslagen doordat de gevonden covariantiematrix, gemeten naar normaliteitsmaatstaven, erg ver van de populatiecovariantiematrix komt te liggen. Men zou dan te snel tot verwerping van een juiste hypothese van m factoren kunnen komen.

Een andere onaangename situatie doet zich voor als in principe wel een factormodel met normaal verdeelde variabelen geldt, maar een klein aantal waarnemingen "besmet" is met grove fouten, zogenaamde uitschieters. Bekend is dat schatters van correlatiecoëfficiënten daarvoor zeer gevoelig zijn en daardoor de schatters in de factoranalyse eveneens. Merk op dat het opvullen van gaten in het waarnemingsmateriaal, als er gedeeltelijk ontbrekende waarden zijn, d.m.v. een of ander procédé (bijvoorbeeld invullen van het gemiddelde van de wel aanwezige waarden) eveneens pseudowaarnemingen creëert, die niet volgens het model gegenereerd zijn, met onbekende gevolgen.

Men zou kunnen vermoeden dat de geschatte standaardafwijkingen van de schatters redelijk zullen voldoen in geval van niet-normaliteit indien normaliteitstoetsen, gebaseerd op derde en vierde moment van de variabelen, niet verwerpen.

In 3.7 wordt ingegaan op wat bekend is als geen normaliteit verondersteld wordt. Het komt erop neer dat men ook dan schattingen kan verkrijgen die met willekeurig grote kans willekeurig dicht bij de werkelijke parameters liggen, mits men maar voldoende veel waarnemingen doet. Helaas is onbekend hoeveel voldoende veel is.

## 2.12. HET GUTTMAN CRITERIUM

Na ontdekt te hebben dat een m-factormodel met zekere ladingenmatrix bij de waarnemingen past, wil men veelal een betekenis toekennen aan de zo gevonden factoren. Of men wil factorscores uitrekenen, d.w.z. de waarden schatten die de factoren aannemen bij de diverse experimentele eenheden.

Beide voornemens kunnen lang niet altijd succesvol worden vervuld: want zelfs als de ontbinding $\Sigma = \Lambda\Lambda' + \Psi$ bij een vast gekozen rotatie uniek is, zijn de factoren $\underline{\phi}_1,\ldots,\underline{\phi}_m$ niet uniek, niet identificeerbaar. Dit is het factoridentificatieprobleem.

Het enige wat wij op grond van de analyse van de factoren weten is:

a. de covariantie van factor $\underline{\phi}_j$ met variabele $\underline{x}_i$ is $\lambda_{ij}$ (in feite beschikken we zelfs niet over deze kennis, maar slechts over een schatting van deze factorlading)
b. de factoren zijn onderling ongecorreleerd, hebben variantie 1 en zijn ongecorreleerd met de specifieke gedeelten, die onderling eveneens ongecorreleerd zijn.

Dit bepaalt de factoren niet volledig. Er blijken nog zeer vele variabelen te bestaan die alle aan deze twee condities voldoen en dus alle met evenveel recht als factoren kunnen worden beschouwd. We zullen ze kandidaatfactoren noemen.

Laten $(\underline{\phi}_1,\ldots,\underline{\phi}_m)$ en $(\underline{\phi}^*_1,\ldots,\underline{\phi}^*_m)$ twee rijen kandidaat factoren bij dezelfde factorladingenmatrix voorstellen en laten $(\underline{\varepsilon}_1,\ldots,\underline{\varepsilon}_p)$ en $(\underline{\varepsilon}^*_1,\ldots,\underline{\varepsilon}^*_p)$ de bijbehorende rijtjes van specifieke gedeelten zijn.

Dan geldt dus

$$\underline{x}_i = \sum_{j=i}^{m} \lambda_{ij} \underline{\phi}_j + \underline{\varepsilon}_i = \sum_{j=i}^{m} \lambda_{ij} \underline{\phi}_j^* + \underline{\varepsilon}_i^*, \quad i = 1,\ldots,p$$

of, in termen van de aan de experimentele eenheden te observeren waarden

$$\underline{x}_{ik} = \sum_{j=i}^{m} \lambda_{ij} \underline{\phi}_{jk} + \underline{\varepsilon}_{ik} = \sum_{j=i}^{m} \lambda_{ij} \underline{\phi}_{jk}^* + \underline{\varepsilon}_{ik}^*,$$

$$i = 1,\ldots,p; \; k = 1,\ldots,n.$$

Zij $\rho_j$ de kleinste waarde van de correlatiecoëfficiënt tussen $\underline{\phi}_j$ en $\underline{\phi}_j^*$, als $\underline{\phi}_j$ en $\underline{\phi}_j^*$ alle mogelijke tweetallen kandidaten voor dezelfde factor zijn.

Wij noemen $\rho_j$ de *Guttman criteriumwaarde* (GCW) voor de j-de factor. (GUTTMAN (1955) heeft als eerste het factoridentificatieprobleem aan de orde gesteld.)

Men kan laten zien dat er voor elke waarde $\rho$ met $\rho_j \leq \rho \leq 1$ kandidaatfactoren $\underline{\phi}_j$ en $\underline{\phi}_j^*$ bestaan zodanig dat

$$\rho(\underline{\phi}_j, \underline{\phi}_j^*) = \rho$$

Voor een constructie van dergelijke variabelen, zie 3.5.

Wanneer de GCW $\rho_j$ dicht bij 1 ligt, lijken alle kandidaten voor de j-de factor sprekend op elkaar en men kan deze factor veilig interpreteren en eventueel factorscores schatten. Maar als $\rho_j$ klein is -en Guttman en anderen hebben waarden kleiner dan nul gevonden in vele gepubliceerde en, wat erger is, geïnterpreteerde gevallen- kunnen noch de ladingenmatrix $\Lambda$, noch de factorscores veel zeggen over de onderliggende factor: men kan dan immers een radicaal verschillende variabele met gelijke kracht verdedigen. De "factor" is zo vaag dat interpretaties die een zeer geringe correlatie hebben te verdedigen zijn.

Het is dan ook geen wonden dat, hoewel er vele methoden bestaan om factorscores te schatten, geen methode echt kan voldoen omdat het benaderde concept zelf zo vaag is: een factorscore moet namelijk de score op een hele familie van kandidaatfactoren representeren.

Met het *Guttman criterium* bedoelen we nu de regel:
een factor kan alleen dan met recht worden geïnterpreteerd als zijn Guttman-criteriumwaarde een zekere minimale waarde overschrijdt.

Hoe groot die minimale waarde dan moet zijn is grotendeels een kwestie van smaak. Het wordt natuurlijk beïnvloed door wat men verder met de factoren gaat doen. De schrijvers dezes houden 0,75 als minimum aan.

Het is vooral wegens het factoridentificatieprobleem dat factoranalyse minder geschikt is voor datareductie. Ook al past een factormodel, dan is nog niet gegarandeerd dat factorscores een bruikbare samenvatting van het materiaal geven omdat ze zo slecht bepaald zijn. Principale componentenanalyse mag ook zijn onduidelijkheden hebben, het is daarbij tenminste duidelijk wat een component is.

Meestal is het zo dat bij een factormodel met verscheidene factoren er een aantal zijn met hoge GCW's en andere met te lage, zodat alleen de eerste voor verdere doeleinden kunnen worden gebruikt.

Voor de berekening van de GCW's verwijzen we naar 3.5. We merken op dat het berekenen van GCW's helaas geen gewoonte is: de meeste computerprogramma's laten het achterwege. De programma's in STATAL (1976) voorzien er wel in.

Tenslotte vermelden we dat de GCW's niet ongevoelig zijn voor rotatie. De bespreking van dit verschijnsel is opgenomen in 2.13 omdat daar rotatie in het algemeen besproken wordt. Hier vermelden we slechts dat het noodzakelijk is de GCW's na een rotatie opnieuw te berekenen en dat GCW's voor een zogenaamde scheve rotatie moeilijk te interpreteren zijn. Ze geven dan vaak een te optimistisch beeld.

## 2.13. INTERPRETATIE EN ROTATIE

Als $\Sigma = \Lambda\Lambda' + \Psi$ en $\Theta$ is een $(m \times m)$-orthogonale matrix (i.e. $\Theta\Theta' = I$) dan geldt ook $\Sigma = \Lambda\Theta\Theta'\Lambda' + \Psi = (\Lambda\Theta)(\Lambda\Theta)' + \Psi$, d.w.z. ook $(m, \Lambda\Theta, \Psi)$ is een factoroplossing voor $\Sigma$, voor elke orthogonale $\Theta$. Vertalen we dit in termen van factoren dan geldt naast

$$\underline{x} = \Lambda\underline{\phi} + \underline{\varepsilon}$$

tevens

$$\underline{x} = \Lambda\Theta\Theta'\underline{\phi} + \underline{\varepsilon} \qquad \text{(met } \textit{dezelfde } \underline{\varepsilon}\text{)}$$

zodat naast

$$\underline{\phi} = (\underline{\phi}_1, \ldots, \underline{\phi}_m)' \qquad \text{(met factorladingen } \Lambda\text{)}$$

ook

$$\Theta'\underline{\phi} \qquad \text{(met factorladingen } \Lambda\Theta\text{)}$$

als rijtje factoren kan optreden.

Wij weten, op dit punt gekomen, dat het wiskundig gezien niet onmogelijk is de variabelen $\underline{x}$ te denken als voortgebracht door een m-tal factoren en een p-tal specifieke gedeelten.

Interpretatie is: het geven van een betekenis aan gepostuleerde factoren èn het doen van de uitspraak dat deze factoren in werkelijkheid bestaan en dat de variabelen $\underline{x}$ er inderdaad zo van afhangen. (Vgl. 1.12.)

Nu kan de situatie zich voordoen dat men het niet mogelijk acht de gevonden $\underline{\phi}$ op deze wijze te interpreteren. Men kan zich dan afvragen of door bepaalde keuze van een orthogonale $\Theta$, $\Theta'\underline{\phi}$ wel verantwoord geïnterpreteerd kan worden.

Aangezien er oneindig veel orthogonale $\Theta$ bestaan, is het niet mogelijk ze alle te beschouwen. Daarom zijn er vuistregels in omloop die in de praktijk de interpretatie bleken te bevorderen.

Het is mogelijk stochastische variabelen voor te stellen als elementen van een lineaire ruimte met de covariantie als inproduct. Grafisch betekent dit dat men gestandaardiseerde variabelen als eenheidsvectoren kan voorstellen, die een hoek met elkaar maken, waarvan de cosinus gelijk is aan de correlatiecoëfficiënt tussen deze variabelen. Wanneer er slechts twee factoren zijn, kan men een plaatje tekenen van deze factoren (fig. 2.13.1).

Het weergeven van de originele variabelen is niet mogelijk, maar dit het tekenen van p nieuwe dimensies voor de specifieke gedeelten vereist.

Wel kunnen de projecties van de variabelen in het vlak van $\phi_1$ en $\phi_2$ getekend worden. Nu is in zo'n plaatje $\Theta'\underline{\phi}$ te zien als een draaiïng van het assenkruis (vandaar de naam rotatie), fig. 2.13.2.

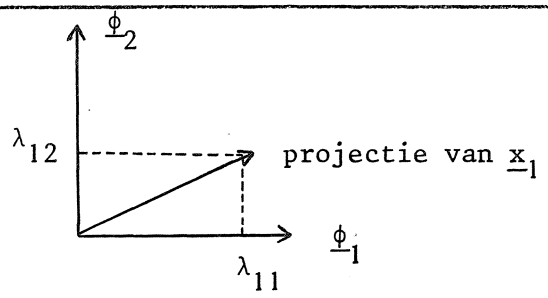



fig. 2.13.1

Grafische voorstelling van een variabele op 2 factoren

Een vuistregel van Thurstone is nu: interpretatie wordt gemakkelijker als men zo'n rotatie kiest dat: veel punten (i.e. projecties van variabelen in het vlak der factoren) dicht bij één van de geroteerde assen liggen terwijl weinig punten ver verwijderd zijn van alle assen.

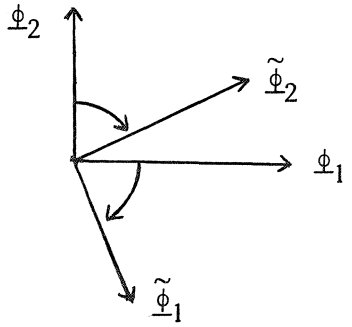



fig. 2.13.2

Een orthogonale rotatie

Het idee is dat factoren, die met een beperkt aantal variabelen flink correleren gemakkelijk te herkennen zijn als een bepaalde betekenis hebbend, terwijl overwegingen van eenvoud ingeven dat de meeste variabelen van slechts één factor zouden behoren af te hangen.

Er kan niet genoeg op gewezen worden, dat zo'n rotatie geen garantie voor succesvolle interpretatie is. De verantwoordelijkheid om een zekere rotatie te interpreteren ligt geheel bij de onderzoeker.

VOORBEELD

| $\Lambda$: | $\underline{\phi}_1$ | $\underline{\phi}_2$ |
|---|---|---|
| $\underline{x}_1$ | 0,75 | 0,63 |
| $\underline{x}_2$ | 0,69 | 0,57 |
| $\underline{x}_3$ | 0,80 | 0,49 |
| $\underline{x}_4$ | 0,85 | -0,42 |
| $\underline{x}_5$ | 0,76 | -0,42 |

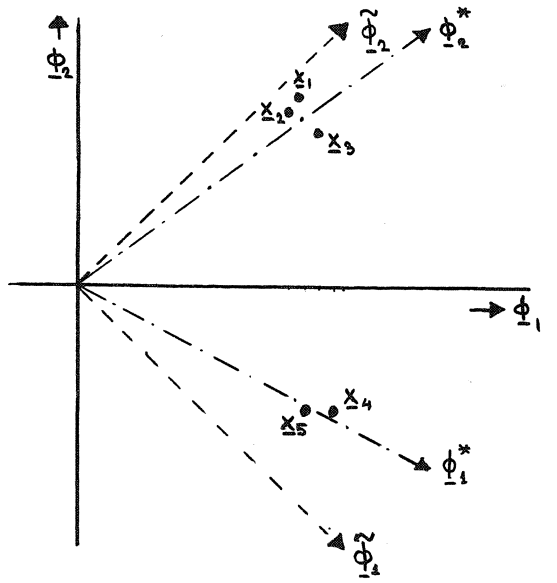



fig. 2.13.3
Het zoeken van een geschikte rotatie

Men kan als geroteerd assenkruis bijvoorbeeld $\tilde{\underline{\phi}}_1$, $\tilde{\underline{\phi}}_2$ kiezen om zoveel mogelijk aan Thurstone's criterium te voldoen. Dit leidt dan tot

| $\tilde{\Lambda}$: | $\tilde{\underline{\phi}}_1$ | $\tilde{\underline{\phi}}_2$ |
|---|---|---|
| $\underline{x}_1$ | 0,14 | 0,95 |
| $\underline{x}_2$ | 0,14 | 0,90 |
| $\underline{x}_3$ | 0,18 | 0,92 |
| $\underline{x}_4$ | 0,94 | 0,09 |
| $\underline{x}_5$ | 0,92 | 0,07 |

en het hangt van het onderwerp in kwestie af of dit beter te interpreteren is.

Sommige onderzoekers hopen interpretatie te vergemakkelijken door ook scheefhoekige assenstelsels te beschouwen, hetgeen betekent dat de geroteerde factoren $\tilde{\underline{\phi}}_1,\ldots,\tilde{\underline{\phi}}_m$ niet ongecorreleerd zijn. Men spreekt van scheve rotaties. Zij voeren aan dat er geen principiële redenen zijn om te eisen dat factoren ongecorreleerd zijn, maar dat vaak a priori aanwezig geachte factoren volgens de theorie waarbinnen gewerkt wordt, gecorreleerd zullen zijn en dat scheefhoekige assenstelsels in het algemeen Thurstone's vuistregel beter kunnen naleven (zie fig. 2.13.3, de factoren $\underline{\phi}_1^*$ en $\underline{\phi}_2^*$). Of ze ook gemakkelijker tot interpretatie leiden valt nog te bezien. Dit blijft

ter beoordeling aan de onderzoeker zelf, die daarin tevens de rijmbaarheid van de correlatiecoëfficiënten tussen de factoren dient te betrekken, binnen de theorie waarin wordt gewerkt.

Wanneer het aantal factoren groter is dan twee, kan men niet het hele rotatieproces in een plaatje vangen. Daarom zijn procedures ontwikkeld, die op niet-grafische manier interpretatie trachten te bevorderen, alle op basis van Thurstone's vuistregel.

Voor al deze zogenaamde analytische rotatiemethoden geldt weer, dat ze geen goede resultaten garanderen, maar hopen een interpretatie naderbij te brengen. Er zijn altijd wel gevallen waarin ze daarin niet slagen. Het bestuderen van plaatjes blijft hierbij dan ook nuttig, om te zien, of de methoden een beetje geslaagd zijn in het nakomen van Thurstone's adagium. Wij wijzen er met nadruk op, dat men geheel vrij is om een door zo'n methode geproduceerde rotatie zelf naar eigen inzicht verder te roteren. Uiteindelijk komt het op het inzicht van de onderzoeker aan.

Er bestaan talloze analytische rotatiemethoden.
Tot de meest populaire behoren onder de orthogonale rotatiemethoden

a. De *Varimaxmethode* (HARMAN (1960)), wel de meest populaire. Deze richt zich vooral op het vinden van factoren die met een beperkt aantal variabelen $\underline{x}_i$ gecorreleerd zijn.
b. De *Quartimaxmethode* (HARMAN (1960)) richt zich meer op het vinden van zulke factoren, dat elke variabele van zo weinig mogelijk factoren flink afhankelijk is.
c. De *Equimaxmethode* (NIE e.a. (1975)) tracht beide doeleinden tegelijk te verwezenlijken.

Van al deze methoden bestaan weer twee varianten: de gewone methode en de genormaliseerde methode, welke laatste gelijk is aan het toepassen van de gewone methode op $(\text{diag}\ (\Lambda\Lambda'))^{-\frac{1}{2}}\Lambda$ in plaats van op $\Lambda$ zelf. Dit heeft tot gevolg, dat ook variabelen met lage communaliteiten meetellen bij het eindresultaat van rotatie, hetgeen bij de gewone methode minder het geval is. De genormaliseerde methoden zijn iets gangbaarder.

Onder de scheefhoekige rotatiemethoden behoren de promaxmethode (HENDRICKSON & WHITE (1969)), de oblimax-, de quartimin- en de oblimin-methode tot de populairste (HARMAN (1960)). Deze methoden hebben als na-

deel dat de mate van scheefhoekigheid op erg ondoorzichtige wijze schijnt te worden bepaald. Na een scheefhoekige rotatie vormt de ladingenmatrix niet meer de matrix van covarianties of correlatiecoëfficiënten tussen variabelen en factoren. Voor de interpretatie heeft men die wel nodig. De meeste programmatuur voorziet daar ook wel in.

Tenslotte moet men na een rotatie te hebben uitgevoerd, zich opnieuw bezighouden met de nauwkeurigheid van de schattingen voor de geroteerde ladingen (zie 1.10) en met de GCW's van de geroteerde factoren (zie 1.11 en 2.12), daar deze kenmerken van de oplossing niet ongevoelig zijn voor rotatie.

Over Guttman criterium en rotatie valt nog het volgende te zeggen: wanneer de niet geroteerde oplossing de zogenaamde canonieke oplossing is (d.w.z. $\Lambda'\Psi^{-1}\Lambda$ is een diagonaalmatrix), zoals bij de meeste programmatuur het geval is, dan beschikken we in zekere zin over de best mogelijke GCW's onder alle mogelijke rotaties, d.w.z. dat rotatie de niet-geïdentificeerdheid van factoren alleen maar verergert.

Precieser geformuleerd: de canonieke oplossing is zodanig dat de GCW van de eerste factor maximaal is; dat, onder conditie dat de eerste factor door deze eis is gebonden, de tweede factor een maximale GCW heeft, enz.

Dit houdt in, dat rotatie de GCW van de eerste factor verlaagt en voor latere factoren iets kan verhogen ten koste van slechtere geïdentificeerdheid van eerdere factoren. (Hierbij laten we de *volgorde* van factoren bepalen door de hoogte van hun GCW.) Het is dus zeker zaak de GCW voor de geroteerde oplossing opnieuw te bekijken.

De GCW's van scheef geroteerde factoren zijn soms nogal misleidend optimistisch over de latere factoren. Men kan namelijk door alle factoren een voldoend hoge correlatie met de best gedetermineerde ongeroteerde factor te geven GCW's verkrijgen die willekeurig dicht bij deze hoogste GCW uit de ongeroteerde oplossing liggen. Ze zijn inderdaad dan ook zo goed gedetermineerd, maar dit is van betrekkelijke waarde. Dergelijke factoren zijn goed gedetermineerd *voorzover* ze met de eerste ongeroteerde factor gecorreleerd zijn, terwijl juist het gedeelte waarin ze afwijken van deze factor, dus dat iets *nieuws* beschrijft, niet goed gedetermineerd hoeft te zijn.

In dergelijke gevallen is het raadzaam de GCW's van de canonieke oplossing mede in overweging te blijven nemen en hoge GCW's van fors gecorreleerde factoren met gepaste voorzichtigheid te behandelen.

Over de nauwkeurigheid van schattingen en rotatie valt op te merken, dat niet alle rotatiemethoden een evaluatie van deze nauwkeurigheid toelaten, i.h.b. de grafische methode niet, terwijl de genormaliseerde versie van de hier genoemde analytische rotaties, alsmede de scheefhoekige rotaties, tot vrijwel onhanteerbare algebra aanleiding geven, zie 3.2. In voorkomende gevallen zal men moeten volstaan met bestudering van de nauwkeurigheid van de ongeroteerde (canonieke) oplossing en er verder maar het beste van hopen.

## 2.14. EEN VOORBEELD

Ter illustratie bespreken we aan de hand van het schema (zie 1.0) een bekend voorbeeld uit de factoranalytische literatuur en wel Emmett's heranalyse van data van Slater en Bennett betreffende de aan- of afwezigheid van ruimtelijk inzicht bij 11- en 12-jarige kinderen. (EMMETT (1949), SLATER & BENNETT (1943)) Dit voorbeeld wordt onder anderen door JØRESKOG (1967) aangehaald.

Het blijkt overigens dat we naar Emmett's artikel zelf moeten teruggaan om voldoende informatie over dit voorbeeld te kunnen krijgen om het schema te kunnen gebruiken: overnemers van dit voorbeeld zijn uiterst beperkt in het opgeven van kenmerken van Emmett's geval.

Over de achtergrond het volgende: Slater en Bennett hebben gegevens verzameld die volgens hun aangeven dat er geen ruimtelijk inzichtfactor bij 11- en 12-jarige kinderen voorkomt. Emmett wil aantonen, dat er juist wel zo'n factor bestaat en dat dat ook in Slater en Bennett's eigen materiaal naar voren komt.

Hij analyseert de samenhang van 9 variabelen, ontstaan uit pooling van 17 variabelen van Slater en Bennett. Dit gebeurt omdat in 1949 een 17-variabelen maximum likelihoodoplossing, die Emmett als oplossingsmethode kiest, nog niet binnen aanvaardbare tijd kon worden gevonden.

Emmett is niet erg te spreken over de kwaliteit van de gebruikte tests en het is opmerkelijk dat juist dit gewraakte geval een standaard-

voorbeeld werd.

De negen variabelen zijn verdeeld in drie groepen van drie: drie niet-verbale intelligentietestscores (variabelen 1, 2 en 3), drie verbale intelligentietestscores (4, 5 en 6) en drie ruimtelijk inzicht-testscores (7, 8 en 9). De correlatiematrix, gebaseerd op de scores van 211 kinderen is weergegeven in tabel 2.14.1.

| variabele | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| variabele 1 | 1,000 | | | | | | | | |
| 2 | 0,523 | 1,000 | | | | | | | |
| 3 | 0,395 | 0,479 | 1,000 | | | | | | |
| 4 | 0,471 | 0,506 | 0,355 | 1,000 | | | | | |
| 5 | 0,346 | 0,418 | 0,270 | 0,691 | 1,000 | | | | |
| 6 | 0,426 | 0,462 | 0,254 | 0,791 | 0,679 | 1,000 | | | |
| 7 | 0,576 | 0,547 | 0,452 | 0,443 | 0,383 | 0,372 | 1,000 | | |
| 8 | 0,434 | 0,283 | 0,219 | 0,285 | 0,149 | 0,314 | 0,385 | 1,000 | |
| 9 | 0,639 | 0,645 | 0,504 | 0,505 | 0,409 | 0,472 | 0,680 | 0,470 | 1,000 |

tabel 2.14.1
Correlatiematrix van 9 variabelen

Aan de hand van het schema valt het volgende te zeggen.
Het doel van de analyse (vraag 1) is exploratie. Emmett wil nagaan hoeveel factoren nodig zijn om de data te beschrijven, daar hij Slater en Bennett's bewering dat het er 2 zouden zijn aanvechtbaar vindt. Over de aard van de steekproef (vraag 2) heeft Emmett diepgaande twijfel. Slater en Bennett beschouwen hun 211 proefpersonen als een aselecte trekking uit de 11- en 12-jarige schoolkinderen in het jaar van testafname, doch Emmett zou zelf onderscheid gemaakt hebben tussen jongens en meisjes. Hij zou dan twee steekproeven hebben genomen, uit jongens en uit meisjes, en de testscores apart hebben geanalyseerd, omdat hij bij beide groepen een verschillende factorstructuur verwacht. De gegevens van Slater en Bennett laten dit onderscheid niet meer toe.

Wij geloven dat Emmett hier een belangrijk punt aanstipt en dat het wezenlijke verschil van inzicht tussen Emmett enerzijds en Slater en Bennett anderzijds niet zozeer gelegen is in de vraag of de correlatie-

matrix uit tabel 2.14.1 wel of niet een 3-factoroplossing toelaat, maar meer in de vraag of deze matrix wel of niet als basis van onderzoek kan dienen. Zoals in 2.6 uiteengezet, kan het op één hoop gooien van waarnemingen, die uit twee populaties afkomstig zijn, extra factoren introduceren. Evengoed kan het zijn dat een factor, welke in beide populaties op zich aanwezig is in de gecombineerde populatie wegvalt (vgl. fig. 2.14.1: in de populaties "meisjes" en "jongens" apart bestaat een duidelijke samenhang tussen de variabelen $x_1$ en $x_2$; in de gecombineerde populatie kan echter $\text{cov}(\underline{x}_1, \underline{x}_2)$ nul zijn). Verder zijn allerlei tussenvormen mogelijk. Emmett werkt echter door, ondanks zijn twijfel aan de adequaatheid van het waarnemingsmateriaal.

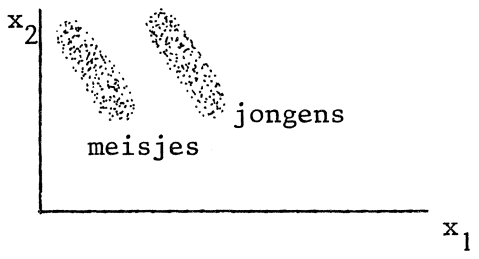



fig. 2.14.1

Wegvallen van samenhang in gecombineerde populatie

Of Emmett zich verdiept heeft in de lineariteit van samenhangen (vraag 3) weten wij niet. Binnen de toenmalige opgeld doende psychologische theorie is het mogelijk te bevestigen dat de variabelen op intervalniveau (vraag 3) zijn gemeten. Emmett kiest voor het algemene model (vraag 4), waardoor werken met de correlatiematrix mogelijk is, omdat AFA immers niet schaalgevoelig is. Uit vraag 5 volgt dat het maximale aantal zinvol te extraheren factoren 5 (=entier $(9+\frac{1}{2}-\frac{1}{2}\sqrt{73})$) is.

We mogen veronderstellen dat Emmett inderdaad heeft nagedacht over identificeerbaarheid van zijn model (vraag 8). Het aantal waarnemingen (vraag 9) is 211 en derhalve veel groter dan tien maal het aantal variabelen. Over multivariate normaliteit (vraag 11) zwijgt Emmett, dit is voor ons niet te achterhalen. Gaan we daarvan uit, dan blijkt de modeltoets een 0-factormodel duidelijk te verwerpen (tabel 2.14.2), zodat we volgens vraag 15 terugspringen naar vraag 4. De antwoorden op vraag 5, 8, 9 en 11 blijven gelijk. Ook een 1-factormodel wordt verworpen en na een

nieuwe cyclus door de vragen geeft een 2-factoroplossing bij vraag 13 een overschrijdingskans van 0,09, zodat de visie van Slater en Bennett dat een 2-factormodel geschikt is weinig verwondering wekt. Emmett besloot echter elke toets bij onbetrouwbaarheid 0,10 uit te voeren en hij probeert een 3-factormodel, dat door de modeltoets niet wordt verworpen. Hij neemt aan dat 3 factoren de samenhang tussen de variabelen kunnen beschrijven. (Emmett gebruikte een maximum likelihoodmethode van Lawley, MLR is de termen van 2.3; wij hebben een heranalyse uitgevoerd met de methode MLJ, volgens Jøreskog. De hier vermelde resultaten zijn op deze laatste methode gebaseerd. Zij verschillen hier en daar enigszins van Emmett's oorspronkelijke schattingen door de iets betere numerieke techniek. Vergelijk ook de heranalyse van deze data van LAWLEY & MAXWELL (1971).)

| m | toetsings-grootheid | df | overschrijdings-kans (gebaseerd op $\chi^2$-benadering) |
|---|---|---|---|
| 0 | 972 | 36 | $< 10^{-4}$ |
| 1 | 230 | 27 | $< 10^{-4}$ |
| 2 | 27,5 | 19 | 0,09 |
| 3 | 7,1 | 12 | 0,85 |

tabel 2.14.2
Modeltoetsen voor Emmett's data

Maximum likelihoodschattingen voor factorladingen en uniciteiten in het 3-factormodel zijn weergegeven in tabel 2.14.3, voor de canonieke oplossing. De geschatte standaardafwijkingen van de ladingenschatters (vraag 14) staan vermeld in tabel 2.14.4. Zij zijn in het algemeen bevredigend klein (vgl. 1.10), zeker voor Emmett's doeleinden: hij wil slechts laten zien dat bij elke van de drie factoren ladingen zijn die significant van nul verschillen.

Anders ligt het bij de Guttman criteriumwaarden (vraag 16): deze zijn vermeld in tabel 2.14.5. De GCW voor de derde factor is negatief zodat deze volstrekt niet te identificeren is, terwijl men ook over de tweede factor niet enthousiast kan zijn.

Emmett was van deze zwakheden van zijn factoren niet op de hoogte en

| variabele | factor I | II | III | $\Psi$ |
|---|---|---|---|---|
| 1 | 0,664 | 0,321 | 0,074 | 0,450 |
| 2 | 0,689 | 0,247 | -0,193 | 0,427 |
| 3 | 0,493 | 0,302 | -0,222 | 0,617 |
| 4 | 0,837 | -0,292 | -0,035 | 0,212 |
| 5 | 0,705 | -0,315 | -0,153 | 0,381 |
| 6 | 0,819 | -0,377 | 0,105 | 0,177 |
| 7 | 0,661 | 0,356 | -0,078 | 0,400 |
| 8 | 0,458 | 0,295 | 0,491 | 0,462 |
| 9 | 0,766 | 0,428 | -0,012 | 0,231 |

tabel 2.14.3
Canonieke factorladingen, en uniciteiten, 3-factormodel

| variabele | factor I | II | III |
|---|---|---|---|
| 1 | 0,066 | 0,058 | 0,076 |
| 2 | 0,064 | 0,061 | 0,068 |
| 3 | 0,070 | 0,071 | 0,083 |
| 4 | 0,060 | 0,046 | 0,045 |
| 5 | 0,065 | 0,057 | 0,072 |
| 6 | 0,064 | 0,046 | 0,037 |
| 7 | 0,068 | 0,057 | 0,066 |
| 8 | 0,073 | 0,093 | 0,142 |
| 9 | 0,066 | 0,047 | 0,055 |

tabel 2.14.4
Standaardafwijkingen van canonieke ladingen uit tabel 2.14.3

trachtte een interpretatie van de factoren te vinden. Hij heeft vastomlijnde ideëen, voortkomend uit zijn psychologische theorie, over de aard van mogelijke factoren. Daarbinnen passen kennelijk de canonieke factoren niet, want zonder discussie gaat hij over tot (orthogonale) rotatie: het valt hem niet moeilijk een rotatie te vinden die met deze ideëen strookt. Na inspectie van de canonieke ladingen besluit hij zo te roteren dat de correlatie tussen variabele 3 en de factoren 1 en 2 nul is. Aan een eveneens mogelijke derde eis schijnt hij geen behoefte te hebben.

| variabele | factor I | II | III |
|---|---|---|---|
| 1 | 0,659 | 0,073 | 0,332 |
| 2 | 0,738 | 0,149 | 0,077 |
| 3 | 0,619 | 0,000 | 0,000 |
| 4 | 0,536 | 0,686 | 0,168 |
| 5 | 0,462 | 0,637 | 0,014 |
| 6 | 0,430 | 0,749 | 0,278 |
| 7 | 0,747 | 0,001 | 0,203 |
| 8 | 0,333 | -0,013 | 0,654 |
| 9 | 0,822 | 0,036 | 0,303 |

tabel 2.14.6 Geroteerde factorladingen

| factor | GCW |
|---|---|
| I | 0,88 |
| II | 0,54 |
| III | -0,08 |

tabel 2.14.5
Guttman criteriumwaarden van de canonieke factoren

Dit leidt tot het resultaat in tabel 2.14.6. Standaardafwijkingen van deze schatters kunnen wij niet berekenen, in verband met het post hoc karakter van de rotatie. (Indien tevoren is besloten de vermelde ladingen nul te maken kan men wel dergelijke standaardafwijkingen schatten, zie bijvoorbeeld JØRESKOG (1967).) De Guttman criteriumwaarden in tabel 2.14.7 maken echter al duidelijk dat het ons onmogelijk is in te stemmen met Emmett's conclusie dat "... the third factor is ... obviously ... the spatial ability factor ...". Verder dan de hypothese dat een oninterpreteerbare derde factor tot de samenhang bijdraagt, durven wij niet te gaan.

| factor | GCW |
|---|---|
| I | 0,67 |
| II | 0,63 |
| III | 0,03 |

tabel 2.14.7
Guttman criteriumwaarden van geroteerde factoren

Merk op dat Emmett werkt vanuit een duidelijke theorie op zijn vakgebied, die het hem mogelijk maakt op een aantal punten duidelijke beslissingen te nemen. Helaas moet worden geconstateerd dat in veel recenter factoranalytisch werk een dergelijke theorie van waaruit gewerkt wordt nagenoeg ontbreekt.

Tot zover hebben wij Emmett gevolgd en hem slechts aangevuld op het gebied van standaardafwijkingen van schatters en van het Guttman criterium. Laten we echter ook even buiten Emmett's opzet treden en nog wat nadere overwegingen omtrent zijn data naar voren brengen.

In de eerste plaats merken we op, dat, waar Emmett zo'n duidelijk beeld heeft van de verwachte factorstructuur, een toetsende factoranalyse (vraag 1.4) i.p.v. een exploratieve voor de hand ligt. Ten overvloede: dit is niet bedoeld als kritiek op Emmett: in 1949 was deze techniek nog niet ontwikkeld. Daar deze benadering in dit rapport niet behandeld is, zullen wij er verder niet op ingaan. Ook daar doemen trouwens wel moeilijkheden op. Binnen de exploratieve analyse blijvend willen we nog twee punten aan de orde stellen: te weten de mogelijkheid van een scheve rotatie en de

claim van de noodzaak een derde factor te veronderstellen.

Studie van een plaatje van de projectie van de variabelen in een ruimte opgespannen door de drie canonieke factoren (vergelijk 2.13) laat zien dat er twee groepen van variabelen zijn: de variabele intelligentiescores (variabelen 4, 5 en 6) en een groep bestaande uit de niet-verbale intelligentiescores (variabelen 1, 2 en 3) met twee van de ruimtelijk inzichtscores (variabelen 7 en 9) terwijl variabele 8, behorend tot de ruimtelijk inzichtscores, apart staat (figuur 2.14.2). Thurstone's vuistregel (2.13) wordt het best nagekomen door gecorreleerde geroteerde factoren die door de middelpunten van die groepjes gaan, respectievelijk door de projectie van variabele 8. Voeren we dit uit, dan zijn de resultaten als in tabel 2.14.8 (scheef geroteerde ladingenmatrix), tabel 2.14.9 (correlatiecoëfficiënten tussen variabelen en scheef geroteerde factoren), tabel 2.14.10 (correlatiematrix van de scheef geroteerde factoren) en tabel 2.14.11 (GCW van de scheef geroteerde factoren).

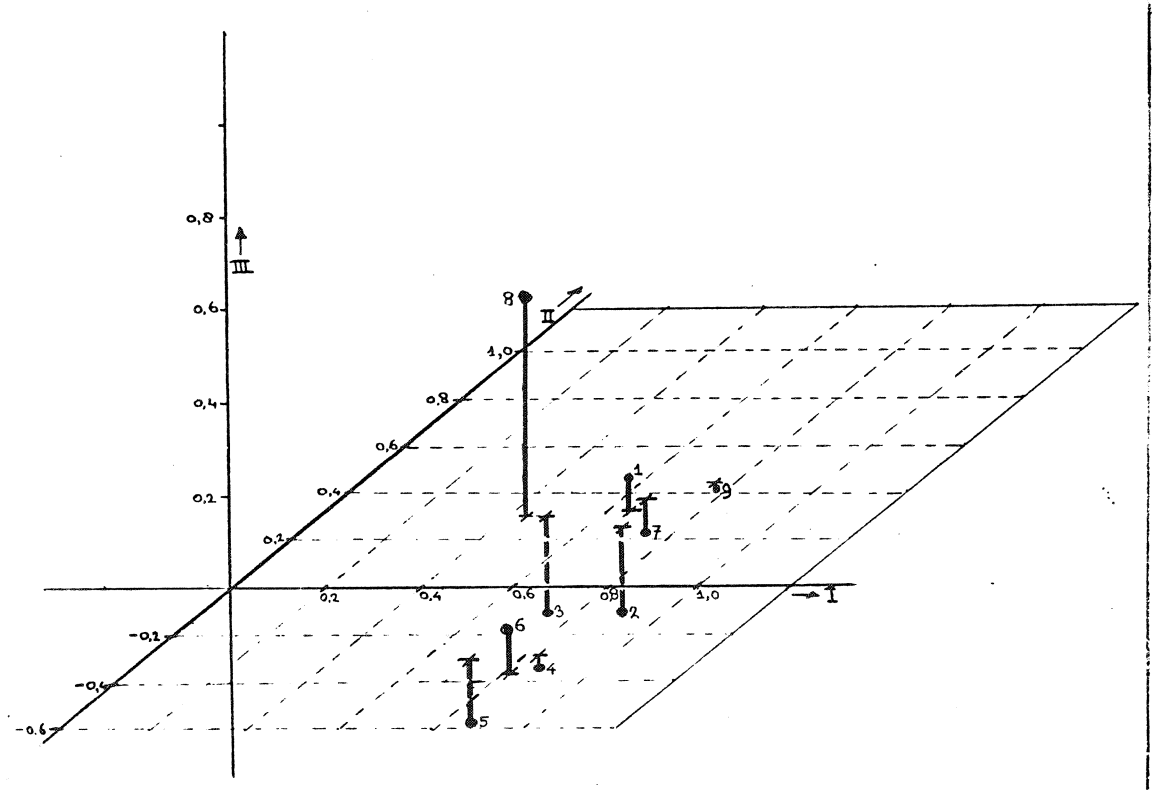



fig. 2.14.2
Schets van de projecties van de variabelen
in de canonieke factorruimte

| variabele | factor I | II | III |
|---|---|---|---|
| 1 | 0,561 | 0,039 | 0,210 |
| 2 | 0,774 | 0,104 | -0,148 |
| 3 | 0,782 | -0,083 | -0,199 |
| 4 | 0,069 | 0,841 | 0,001 |
| 5 | 0,113 | 0,772 | -0,172 |
| 6 | -0,182 | 0,947 | 0,171 |
| 7 | 0,761 | -0,024 | 0,015 |
| 8 | 0,000 | 0,000 | 0,733 |
| 9 | 0,815 | -0,036 | 0,122 |

tabel 2.14.8

Scheef geroteerde ladingenmatrix

| variabele | factor I | II | III |
|---|---|---|---|
| 1 | 0,724 | 0,488 | 0,594 |
| 2 | 0,745 | 0,548 | 0,401 |
| 3 | 0,598 | 0,346 | 0,281 |
| 4 | 0,616 | 0,886 | 0,383 |
| 5 | 0,502 | 0,777 | 0,211 |
| 6 | 0,546 | 0,897 | 0,430 |
| 7 | 0,755 | 0,477 | 0,505 |
| 8 | 0,481 | 0,293 | 0,733 |
| 9 | 0,872 | 0,543 | 0,642 |

tabel 2.14.9

Correlatiecoëfficiënten tussen variabelen en scheef geroteerde factoren

| factor | factor I | II | III |
|---|---|---|---|
| I | 1,000 | | |
| II | 0,650 | 1,000 | |
| III | 0,656 | 0,400 | 1.000 |

tabel 2.14.10

correlatiematrix van de scheef geroteerde factoren

| factor | GCW |
|---|---|
| I | 0,80 |
| II | 0,83 |
| III | 0,39 |

tabel 2.14.11

Guttman criterium-waarden van de scheef geroteerde factoren

Valt deze rotatie te interpreteren? Nu begeven wij ons op glad ijs, als niet-deskundigen op Emmett's vakgebied kunnen wij daar geen beslissend oordeel over geven. Een voorzichtige suggestie onzerzijds is evenwel dat deze testscores geen onderscheid mogelijk maken tussen niet-verbale intelligentie en ruimtelijk inzicht, terwijl variabele 8 verschilt van de andere ruimtelijk inzichtvariabelen. Misschien is er sprake van een niet-lineair verband? Of is het aantal factoren toch te groot gekozen? Het is een vaak geconstateerd verschijnsel dat een overmaat van factoren tot factoren schijnt te leiden die nagenoeg alleen met slechts één variabele gecorreleerd

zijn. Voegen we dit bij de op grond van de modeltoets (tabel 2.14.2) best te verdedigen stellingname van Slater en Bennett dan lijkt het interessant om de analyse te herhalen op het materiaal na weglating van variabele 8. Uitvoering van dit voornemen laat zien dat een 2-factormodel het eerste is dat dan niet wordt verworpen (toetsingsgrootheid = 10.83, asymptotische $\chi^2$ overschrijdingskans met 13 vrijheidsgraden 0,63). Dit resultaat steunt de hypothese dat de derde factor in het volledige probleem alleen aan variabele 8 te wijten is. Bovendien lijkt de ladingenmatrix in het verkleinde probleem vrij goed op de ladingen van de eerste twee factoren (behoudens de ladingen bij de 8-ste variabele) in het volledige probleem, als wederom een scheve rotatie wordt toegepast, zodat de nieuwe factoren door de middelpunten van de groepjes variabelen 1, 2, 3, 7 en 9, respectievelijk 4, 5 en 6 gaan.

De betreffende resultaten zijn vermeld in de tabellen 2.14.12 tot 2.14.15.

| variabele | factor I | factor II | uniciteiten $\Psi$ |
|---|---|---|---|
| 1 | 0,699 | 0,042 | 0,471 |
| 2 | 0,663 | 0,114 | 0,450 |
| 3 | 0,624 | -0,058 | 0,654 |
| 4 | 0,052 | 0,864 | 0,194 |
| 5 | -0,006 | 0,772 | 0,410 |
| 6 | -0,046 | 0,913 | 0,217 |
| 7 | 0,816 | -0,064 | 0,398 |
| 9 | 0,897 | -0,035 | 0,235 |

tabel 2.14.12

Factorladingen en uniciteiten van scheef geroteerde oplossing bij weglaten van variabele no. 8

| variabele | factor I | factor II |
|---|---|---|
| 1 | 0,726 | 0,494 |
| 2 | 0,737 | 0,542 |
| 3 | 0,587 | 0,345 |
| 4 | 0,610 | 0,897 |
| 5 | 0,493 | 0,768 |
| 6 | 0,544 | 0,884 |
| 7 | 0,775 | 0,463 |
| 9 | 0,875 | 0,545 |

tabel 2.14.13

Correlatiecoëfficiënten tussen variabelen en scheef geroteerde factoren bij weglaten van variabele no. 8

| factor | I | II |
|---|---|---|
| I | 1,000 | |
| II | 0,646 | 1,000 |

tabel 2.14.14
Correlatiematrix van de scheef geroteerde factoren bij weglaten van variabele no. 8

| factor | GCW |
|---|---|
| I | 0,69 |
| II | 0,70 |

tabel 2.14.15
Guttman criteriumwaarde van scheef geroteerde factoren bij weglaten van variabele no. 8

Tenslotte blijkt ook in het verkleinde probleem een mogelijke derde factor niet de beoogde ruimtelijk inzichtfactor te zijn: de derde factor blijkt vrijwel uitsluitend met de derde variabele (uit de groep niet verbale intelligentiescores) gecorreleerd te zijn.
Ook deze nadere overwegingen steunen Emmett's conclusie over het bestaan van een derde factor derhalve allerminst.

# 3. DETAILS

Dit deel bevat een verscheidenheid aan onderwerpen: formele specificatie van de gebruikte modellen, een aantal belangrijke feiten uit moeilijk verkrijgbare literatuur, nieuwe resultaten en uiteenzettingen die door hun uitgebreidheid of wiskundige diepgang het betoog in de LEIDRAAD of de UITLEG gestoord zouden hebben. Ook in dit deel geldt dat de onderdelen los van elkaar staan. Hun volgorde heeft dan ook geen betekenis.

## 3.1. DE GEBRUIKTE MODELLEN EN HUN OPLOSSINGSMETHODEN

Dit onderdeel bevat een formele specificatie van de drie in dit rapport centraal staande factormodellen en de daarbij behorende oplossingsmethoden. Een minder technische uiteenzetting staat in 2.1 en 2.3.

*Algemene model* (AFA)

Uitgaande van ongecorreleerde factoren met variantie 1 geldt, als direct gevolg van het algemene model,

$$\Sigma = \Lambda\Lambda' + \Psi$$

met $\Sigma$ de $(p \times p)$-covariantiematrix, niet singulier;
$\Lambda$ een $(p \times m)$-matrix van de rang m;
$\Psi$ een $(p \times p)$-diaginaal matrix met niet-negatieve diagonaalelementen.

Als de factoren covariantiematrix A hebben, wordt deze formule $\Sigma = \Lambda A\Lambda' + \Psi$; tenzij anders vermeld, nemen wij altijd aan dat $A = I$. Voor het geval dat één of meer der diagonaalelementen van $\Psi$ nul zijn, zie 3.6.

De oplossingsmethode bij gegeven m d.m.v. maximum likelihoodschatting staat duidelijk beschreven in LAWLEY & MAXWELL (1971), daarom volgt hier slechts een globale beschrijving. Voor het schatten van m zie 2.10.

Uitgaande van de steekproefcovariantiematrix $\underline{S}$ gebaseerd op n waarnemingen (voor definitie zie 2.2), die een zuivere schatter van $\Sigma$ is, worden, onder aanname van multivariate normaliteit van de waarnemingen maximum likelihoodschatters $\underline{L}$ voor $\Lambda$ en $\underline{P}$ voor $\Psi$ berekend. De likelihood-

functie $\ell$ is afgeleid van de Wishartverdeling van $(n - 1)\underline{S}$ met parameters $n - 1$ en $\Sigma$:

$$\log(\underline{\ell}(\Lambda,\Psi)) = \underline{c} - \tfrac{1}{2}(n - 1)(\log|\Sigma| + \mathrm{tr}(\underline{S}\Sigma^{-1})),$$

waarbij $\Sigma = \Lambda\Lambda' + \Psi$. De matrices $\underline{L}$ en $\underline{P}$ die deze functie maximaliseren zijn de gezochte schatters. In plaats van het maximum van deze functie wordt nu het minimum bepaald van

$$\underline{F}(\Lambda,\Psi) = \log|\Sigma| + \mathrm{tr}(\underline{S}\Sigma^{-1}) - \log|\underline{S}| - p.$$

Dit gebeurt in 2 stappen:

a. Zoek eerst, gegeven $\Psi$, het minimum van $\underline{F}$ en noem dit minimum $\underline{f}(\Psi)$. Het minimum wordt gevonden voor $\Lambda = \underline{L}$, waarbij de kolommen van $\Psi^{-\frac{1}{2}}\underline{L}$ gelijk zijn aan de eigenvectoren van $\Psi^{-\frac{1}{2}}\underline{S}\Psi^{-\frac{1}{2}}$.
b. Minimaliseer $\underline{f}(\Psi)$ over $\Psi$. Men kan bewijzen dat $\underline{f}(\Psi)$ wordt geminimaliseerd wanneer geldt $\Psi = \mathrm{diag}(\underline{S} - \underline{LL}')$, maar het beste oplossingsalgoritme maakt hiervan geen gebruik. Dit berust op numerieke minimalisatie van $\underline{f}$ d.m.v. een gemodificeerde Fletcher- en Powellmethode.

De toetsingsgrootheid om te toetsen of het model past (zie 2.11) is gebaseerd op het likelihoodratiobeginsel en is evenredig met het minimum van $\underline{f}(\Psi)$. Om de asymptotische benadering te verbeteren wordt deze grootheid gemodificeerd met een Box-Bartlettvermenigvuldigingsfactor. De gebruikte toetsingsgrootheid is dan

$$\underline{G} \overset{\text{def}}{=} (n - 1 - \frac{2p + 5}{6} - \frac{2m}{3})\underline{f}(\underline{P})$$

en heeft, wanneer het model geldt èn de normaliteitsassumptie geldt, bij benadering een chi-kwadraatverdeling met

$$\tfrac{1}{2}((p - m)^2 - (p + m)) \quad \text{vrijheidsgraden.}$$

Deze benadering is betrouwbaar als $n \geq p + 50$. Voor meer details zie LAWLEY & MAXWELL (1971).

*Gelijke residuele variantiesmodel* (GFA)

Het gelijke residuele variantiesmodel luidt:

$$\Sigma = \Lambda\Lambda' + \sigma^2 I$$

met $\Sigma$ de (p × p)-covariantiematrix, niet singulier

$\Lambda$ een (p × m)-matrix van de rang m;
$\sigma^2$ een reëel getal > 0;
I de (p × p)-identiteitsmatrix.

Omdat de oplossingsmethode van dit model in de literatuur niet gemakkelijk toegankelijk is, gaan we er hier wat uitgebreider op in.

Notatie: $\Xi \stackrel{def}{=} \Lambda\Lambda'$.

Daar $\Lambda$ van de rang m is, is ook $\Xi$ van de rang m. Dit laatste houdt in, dat $\Xi$ m positieve eigenwaarden $\delta_1,\ldots,\delta_m$ heeft en p - m eigenwaarden gelijk aan 0. Laten $\omega_1,\ldots,\omega_p$ de bijbehorende eigenvectoren zijn. Dan heeft $\Sigma$ eigenwaarden $\delta_1 + \sigma^2,\ldots,\delta_m + \sigma^2,\sigma^2,\ldots,\sigma^2$ met dezelfde eigenvectoren $\omega_1,\ldots,\omega_p$.
Immers als $i \leq m$ dan:
$\Sigma\omega_i = (\Xi + \sigma^2 I)\omega_i = \delta_i\omega_i + \sigma^2\omega_i = (\delta_i + \sigma^2)\omega_i$;
en als $i > m$ dan:
$\Sigma\omega_i = (\Xi + \sigma^2 I)\omega_i = 0 + \sigma^2\omega_i = \sigma^2\omega_i$

Notatie:

$$\Omega = \begin{bmatrix} \omega_{11} & \cdots & \omega_{m1} \\ \vdots & & \vdots \\ \omega_{1p} & \cdots & \omega_{mp} \end{bmatrix}$$

is de matrix met als kolommen de genormeerde eigenvectoren $\omega_1,\ldots,\omega_m$

$$\Delta = \begin{bmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_m \end{bmatrix}$$

is de (m × m)-diagonaal matrix met als diagonaalelementen de eigenwaarden $\delta_1,\ldots,\delta_m$

Nu geldt (zie bijv. MORRISON (1976)):

$$\Xi = \Omega\Delta\Omega'$$

Definieer nu:

$$\Lambda_0 \stackrel{def}{=} \Omega\Delta^{\frac{1}{2}}$$

Hieruit volgt:

$$\begin{aligned}\Sigma &= \Xi + \sigma^2 I = \Omega\Delta\Omega' + \sigma^2 I = \\ &= (\Omega\Delta^{\frac{1}{2}})(\Omega\Delta^{\frac{1}{2}})' + \sigma^2 I = \\ &= \Lambda_0\Lambda_0' + \sigma^2 I.\end{aligned}$$

Dus $\Lambda_0$ is een oplossing van het GFA-model. Zoals elders uiteengezet, is dan ook $\Lambda_0\Theta$ voor elke orthogonale $\Theta$ een oplossing. Om cumulatie van indices te voorkomen schrijven we verder $\Lambda$ i.p.v. $\Lambda_0$.

We kunnen dus $\Lambda$ uitrekenen als we de eerste m eigenwaarden en eigenvectoren van $\Xi$ weten. We kennen echter $\Xi$ helemaal niet. Gelukkig heeft $\Sigma$ dezelfde eigenvectoren als $\Xi$ en de eerste m eigenwaarden van $\Sigma$ zijn $\sigma^2$ groter dan die van $\Xi$. $\sigma^2$ is bekend daar de laatste p - m eigenwaarden van $\Sigma$ gelijk zijn aan $\sigma^2$.

Om tot een goede interpretatie van de factoren te kunnen komen, wordt gekeken naar de correlatiecoëfficiënten tussen factoren en variabelen:

$$\begin{aligned}\mathrm{cov}(\underline{x}_i, \underline{\phi}_j) &= \mathrm{cov}\Big(\sum_{k=1}^{m} \lambda_{ij}\underline{\phi}_k + \underline{e}_i, \underline{\phi}_j\Big) = \\ &= \mathrm{cov}\Big(\sum_{k=1}^{m} \lambda_{ik}\underline{\phi}_k, \underline{\phi}_j\Big) = \\ &= \lambda_{ij} = \omega_{ij}\cdot\sqrt{\delta_j} \\ \rho(\underline{x}_i, \underline{\phi}_j) &= \omega_{ij}\cdot\frac{\sqrt{\delta_j}}{\sigma_{ii}}\end{aligned}$$

(waarin $\sigma_{ii}$ het i-de diagonaal element van $\Sigma$ is).

In de praktijk beschikken we niet over $\Sigma$ maar slechts over een schatting uit de steekproef $\underline{S}$ voor $\Sigma$. Schematisch kunnen we dan het uitvoeren van GFA factoranalyse als volgt weergeven:

a. Nulhypothese:
   $H_0$: Er zijn m factoren $\underline{\phi}_1,\ldots,\underline{\phi}_m$ zodanig, dat
   $\underline{x} = \Lambda\underline{\phi} + \underline{\varepsilon}$ en $\Sigma = \Lambda\Lambda' + \sigma^2 I$
b. Toets: (onder normaliteit van de waarnemingen)
   Toets of de laatste p - m eigenwaarden van $\underline{S}$ gelijk zijn (M.a.w. of de waarde van m goed is.)
c. Schatten:
   Als de toets niet verwerpt, schat dan $\Lambda$ en $\sigma^2$

Notatie: Laten $\underline{d}_1,\ldots,\underline{d}_p$ de in aflopende grootte geordende eigenwaarden van $\underline{S}$ zijn en $\underline{w}_1,\ldots,\underline{w}_p$ de bijbehorende vectoren.

ANDERSON (1963) heeft de likelihood ratiotoets geconstrueerd om na te gaan of de laatste p - m eigenwaarden van een covariantiematrix gelijk zijn onder normaliteit van de waarnemingen:

$$\underline{G} \overset{\text{def}}{=} \left[ (n-1)(p-m)\left(\log\left(\sum_{i=m+1}^{p} \underline{d}_i\right) - \log(p-m)\right) - \sum_{i=m+1}^{p} \log \underline{d}_i \right]$$

$\underline{G}$ heeft onder de nulhypothese bij benadering voor grote n een chi-kwadraat verdeling met $\frac{1}{2}(p-m)(p-m+1) - 1$ vrijheidsgraden.

Verwerpt de toets, dan is daarmee de geldigheid van het GFA-model, althans voor deze waarde van m, ten gunste van grotere m verworpen.

De volgende stap is, als $H_0$ niet verworpen is, nu het schatten van $\Lambda$ en $\sigma^2$. We zullen schatters $\underline{L}$ en $\underline{s}^2$ bepalen door middel van de maximum-likelihoodmethode onder de veronderstelling dat de waarnemingen multinormaal verdeeld zijn.

Stel, dat $\underline{S}$, de steekproefcovariantiematrix, de waarde S heeft aangenomen. Omdat $(n-1)\underline{S}$ een Wishartverdeling heeft, wordt de likelihoodfunctie $\ell$, gegeven door

$$\log\ell(\Lambda, \sigma^2) = \text{const} - \tfrac{1}{2}(n-1)(\log|\Sigma| + \text{spoor}\,(S\Sigma^{-1})), \text{ waarbij}$$

$$\Sigma = \Lambda\Lambda' + \sigma^2 I.$$

Wij gaan die waarden L en $s^2$ van $\Lambda$ en $\sigma^2$ bepalen die $\ell$ maximaliseren. Maximaliseren van $\ell(\Lambda,\sigma^2)$ komt op hetzelfde neer als minimaliseren van

$$F(\Lambda, \sigma^2) \overset{\text{def}}{=} \log|\Sigma| + \text{spoor}(S\Sigma^{-1})$$

Notatie: $\sigma_{ij}$ is het $i,j^{de}$ element van $\Sigma$
$\sigma^{ij}$ is het $i,j^{de}$ element van $\Sigma^{-1}$.

Matrixdifferentiatie levert op:

$$\frac{\partial \log(|\Sigma|)}{\partial\sigma_{ii}} = \sigma^{ii}; \qquad \frac{\partial \log(|\Sigma|)}{\partial\sigma_{ij}} = 2\sigma^{ij} \quad (i\neq j);$$

$$\frac{\partial \text{ trace}(S\Sigma^{-1})}{\partial\sigma_{ii}} = -(\Sigma^{-1}S\Sigma^{-1})_{ii}; \quad \frac{\partial \text{ trace}(S\Sigma^{-1})}{\partial\sigma_{ij}} = -2(\Sigma^{-1}S\Sigma^{-1})_{ij} \ (i\neq j).$$

Hieruit volgt:

$$\frac{\partial F}{\partial\sigma_{ii}} = (\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1})_{ii}$$

$$\frac{\partial F}{\partial\sigma_{ij}} = 2(\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1})_{ij} \quad (i \neq j)$$

Verder geldt:

$$\sigma_{ii} = \sum_{k=1}^{m} \lambda_{ik}^2 + \sigma^2$$

$$\sigma_{ij} = \sum_{k=1}^{m} \lambda_{ik}\lambda_{jk} \quad (i \neq j)$$

Hieruit volgt:

$$\frac{\partial\sigma_{ii}}{\partial\lambda_{ik}} = 2\lambda_{ik}$$

$$\frac{\partial \sigma_{ij}}{\partial \lambda_{ik}} = \lambda_{jk} \quad (i \neq j)$$

$$\frac{\partial \sigma_{ii}}{\partial \sigma^2} = 1$$

Bovenstaande resultaten gaan we nu toepassen:

$$\frac{\partial F}{\partial \lambda_{ij}} = \sum_{k=1}^{p} \sum_{\ell=1}^{p} \frac{\partial F}{\partial \sigma_{kl}} \frac{\partial \sigma_{k\ell}}{\partial \lambda_{ij}} = \sum_{k=1}^{p} \frac{\partial F}{\partial \sigma_{ik}} \frac{\partial \sigma ik}{\partial \lambda_{ij}} =$$

$$= 2 \sum_{k=1}^{p} (\Sigma^{-1}(\Sigma - S)\Sigma^{-1})_{ik} \lambda_{kj} = 2(\Sigma^{-1}(\Sigma - S)\Sigma^{-1}\Lambda)_{ij}$$

$$\frac{\partial F}{\partial \sigma^2} = \sum_{k=1}^{p} \sum_{\ell=1}^{p} \frac{\partial F}{\partial \sigma_{kl}} \frac{\partial \sigma_{kl}}{\partial \sigma^2} = \sum_{k=1}^{p} \frac{\partial F}{\partial \sigma_{kk}} \frac{\partial \sigma_{kk}}{\partial \sigma^2} =$$

$$= \sum_{k=1}^{p} (\Sigma^{-1}(\Sigma - S)\Sigma^{-1})_{kk}.$$

Conclusie:

$$\frac{\partial F}{\partial \Lambda} = 2\Sigma^{-1}(\Sigma - S)\Sigma^{-1}\Lambda$$

$$\frac{\partial F}{\partial \sigma^2} = \text{spoor}(\Sigma^{-1}(\Sigma - S)\Sigma^{-1}).$$

L en $s^2$ zullen dus moeten voldoen aan:

$$\left(\frac{\partial F}{\partial \Lambda}\right)_{\Lambda = L} = 0; \quad \left(\frac{\partial F}{\partial \sigma^2}\right)_{\sigma^2 = s^2} = 0;$$

Definieer nu:

$$s^2 \stackrel{\text{def}}{=} \left(\sum_{i=m+1}^{p} d_i\right) / (p - m)$$

$$D \stackrel{\text{def}}{=} \begin{bmatrix} d_i - s^2 & & 0 \\ & \ddots & \\ 0 & & d_m - s^2 \end{bmatrix}$$

$$W \stackrel{\text{def}}{=} \begin{bmatrix} w_{11} \cdots w_{m1} \\ \vdots \quad\quad \vdots \\ w_{1p} \cdots w_{mp} \end{bmatrix}, \text{ matrix der eigenvectoren } w_1, \ldots, w_m \text{ van } S$$

$$L \stackrel{def}{=} WD^{\frac{1}{2}}$$

We zullen aantonen dat de aldus gedefinieerde L en $s^2$ aan de voorwaarden voldoen. Het bewijs dat ze ook inderdaad een minimum opleveren wordt achterwege gelaten.
Uit het voorafgaande volgt:

$$SL = SWD^{\frac{1}{2}} = W(D + s^2I)D^{\frac{1}{2}} = WD^{\frac{1}{2}}(D + s^2I) = L(D + s^2I)$$

$$L'L = s^{\frac{1}{2}}W'WD^{\frac{1}{2}} = D.$$

Hieruit volgt:

$$\Sigma L = (LL' + s^2I)L = LL'L + s^2L = LD + s^2L = WD^{\frac{1}{2}}D + s^2WD^{\frac{1}{2}} =$$

$$= W(D + s^2I)D^{\frac{1}{2}} = SL.$$

Dus:

$$\Sigma L = SL = L(D + s^2I) \Rightarrow$$

$$L = \Sigma^{-1}L(D + s^2I) \Rightarrow$$

$$\left(\frac{\partial F}{\partial \Lambda}\right)_{\Lambda=L} = 2\Sigma^{-1}(\Sigma - S)\Sigma^{-1}L = 2\Sigma^{-1}(\Sigma - S)L(D + s^2I)^{-1} =$$

$$= 2\Sigma^{-1}(\Sigma L - SL)(D + s^2I)^{-1} = 0.$$

Als $w_i$ eigenvector is van $S_i$ dan geldt:

$$\Sigma^{-1}w_i = (LL' + s^2I)^{-1}w_i = \frac{1}{d_i}\ w_i$$

$$(\Sigma - S)w_i = (LL' + s^2I - S)w_i = (d_i - s^2 + s^2 - d_i)w_i = 0.w_i$$

Dus: $\Sigma^{-1}(\Sigma - S)\Sigma^{-1}$ heeft eigenvectoren $w_1,\ldots,w_p$ en eigenwaarden 0 $\Rightarrow$
spoor $(\Sigma^{-1}(\Sigma - S)\Sigma^{-1}) = 0 \Rightarrow$

$$\left(\frac{\partial F}{\partial \sigma^2}\right)_{\sigma^2=s^2} = 0.$$

Conclusie (nu niet meer bij de specifieke uitkomst $\underline{S} = S$):

De maximum likelihoodschatters van $\Lambda$ en $\sigma^2$ zijn, bij gegeven m,

$$\underline{L} = \underline{W}\underline{D}^{\frac{1}{2}} \qquad \underline{s}^2 = \frac{1}{p - m} \sum_{i=m+1}^{p} \underline{d}_i;$$

de maximum likelihoodschatter van $\Sigma$ is $\underline{L}\underline{L}' + \underline{s}^2 I$. Hierbij zijn $\underline{d}_1,\ldots,\underline{d}_p$ de eigenwaarden van $\underline{S}$ (in aflopende grootte geordend) en $\underline{w}_1,\ldots,\underline{w}_p$ de bijbehorende eigenvectoren.

$$\underline{D} = \begin{bmatrix} \underline{d}_1 - \underline{s}^2 & & 0 \\ & \ddots & \\ 0 & & \underline{d}_m - \underline{s}^2 \end{bmatrix},$$

$$\underline{W} = \begin{bmatrix} \underline{w}_{11} \cdots \underline{w}_{m1} \\ \vdots \quad\quad \vdots \\ \underline{w}_{1p} \cdots \underline{w}_{mp} \end{bmatrix}, \quad \underline{w}_i = \begin{bmatrix} \underline{w}_{i1} \\ \vdots \\ \underline{w}_{ip} \end{bmatrix}$$

Voor het schatten van m zie 2.10.

Nu is $\underline{S} = \underline{L}\underline{L}' + \underline{P}$, voor zekere $(p \times p)$-matrix $\underline{P}$. Onder de nulhypothese en de veronderstelling dat $\underline{S}$ een goede schatter is voor $\Sigma$, zal nu $\underline{P}$ een matrix moeten zijn die er bij benadering uitziet als $\sigma^2 I$.
De correlatiecoëfficiënten tussen variabelen en factoren schatten we met:

$$\underline{\rho}(\underline{x}_i, \underline{\phi}_j) = \underline{w}_{ij} \cdot \sqrt{\frac{\underline{d}_j}{\underline{S}_{ii}}}$$

Tenslotte kunnen we uitrekenen welk gedeelte van de totale variantie verklaard wordt door de factoren $\underline{\phi}_1,\ldots,\underline{\phi}_m$:

$$\text{percentage verklaarde variantie} = \frac{\sum_{i=1}^{p} \underline{d}_i - ps^2}{\sum_{i=1}^{p} \underline{d}_i} \times 100\ \%.$$

*Jøreskog's model* (JFA)

Jøreskog's model luidt:

$$\Sigma = \Lambda\Lambda' + \theta(\text{diag}(\Sigma^{-1}))^{-1}$$

met $\Sigma$ de $(p \times p)$-covariantiematrix, niet singulier;

$\Lambda$ een $(p \times m)$-matrix van de rang m;

$\theta$ een reëel getal $\in(0,1)$.

De oplossingsmethode d.m.v. kleinste kwadratenschatters staat excellent beschreven in JØRESKOG (1963), daarom hier slechts een globale beschrijving.

Uitgaande van de steekproefcovariantiematrix $\underline{S}$, die een beginschatter voor $\Sigma$ is, worden bij gegeven m schatters $\underline{L}$ voor $\Lambda$ en $\underline{t}$ voor $\theta$ (zonder aanname van multinormaliteit van de waarnemingen) bepaald als die $\Lambda$ en $\theta$ die

$$U(\Lambda,\theta) = \text{spoor}[(\underline{S}^* - \&\text{diag}(\underline{S}^{-1}))^{\frac{1}{2}}\Lambda\Lambda'(\text{diag}(\underline{S}^{-1}))^{\frac{1}{2}} - \theta I)^2]$$

minimaliseren. Hierin is $\underline{S}^* = (\text{diag}(\underline{S}^{-1}))^{\frac{1}{2}}\underline{S}(\text{diag}(\underline{S}^{-1}))^{\frac{1}{2}}$.

Dit leidt tot de volgende schatter voor $\theta$,

$$\underline{t} = \frac{1}{p - m} \sum_{i=m+1}^{p} \underline{c}_i, \text{ waarbij}$$

$\underline{c}_1,\ldots,\underline{c}_p$ de in aflopende grootte geordende eigenwaarden van $\underline{S}^*$ zijn en $\underline{w}_1^*,\ldots,\underline{w}_p^*$ de bijbehorende eigenvectoren, zodanig genormaliseerd dat $\underline{w}_i^{*\prime}\underline{w}_i^* = \underline{c}_i - \underline{t}$.

Dan is de schatter voor $\Lambda$:

$\underline{L} = (\text{diag}(\underline{S}^{-1}))^{-\frac{1}{2}}\underline{W}^*$, waarin $\underline{W}^*$ de matrix met als kolommen deze eigenvectoren $\underline{w}_1^*,\ldots,\underline{w}_m^*$ is.

Voor een modeltoets is wel een assumptie van multivariate normaliteit nodig; de toetsingsgrootheid die Jøreskog voorstelt, is:

$$\underline{G} = n\,(p-m)\log\left(\frac{\sum_{i=m+1}^{p} \underline{c}_i}{p - m}\right) - \log\left(\prod_{i=m+1}^{p} \underline{c}_i\right)\Bigg]$$

Men kan deze toets gebruiken om m te schatten; zie 2.10.

Onder aanname van geldigheid van het model en normaliteit van de waarnemingen, heeft $\underline{G}$ asymptotisch de verdeling van een gewogen som van onafhankelijke chi-kwadraatvariabelen, die volgens Jøreskog in de praktijk kan worden benaderd door een chi-kwadraatverdeling met $\frac{1}{2}(p - m - 1)(p - m + 2)$ vrijheidsgraden.

## 3.2. VARIANTIE VAN DE SCHATTERS

Herhaaldelijk is in dit rapport gewezen op het belang van het beschouwen van de nauwkeurigheid van de gevonden schattingen (zie 1.10 en 2.11) door het berekenen van schattingen van de varianties van de gebruikte schatters.

Daar dit helaas niet tot de routine in de factoranalysepraktijk behoort, zijn deze variantieschatters niet erg bekend, wellicht ook omdat ze vrij ingewikkeld zijn. Daarom vermelden wij hier enige resultaten, en passant ook de covarianties van deze schatters behandelend.

Voor het algemene model volgen wij de presentatie van JENNRICH (1974). Voor het gelijke residuele variantiesmodel zijn de resultaten door ons met zijn methode verkregen. Voor Jøreskog's model volgen wij JØRESKOG (1963). Daarbij gaat het om asymtptotische (co)variantieschatters en in alle gevallen zijn deze (co)varianties functies van de onbekende parameters $\Lambda$ en $\Psi$ (of $\sigma^2$, of $\theta$), zodat in de praktijk schattingen moeten worden verkregen door substitutie van $\underline{L}$ en $\underline{P}$ (resp. $\underline{s}^2$, resp. $\underline{t}$, zie 3.1), hetgeen alleen tot aanvaardbare resultaten leidt voor grote waarden van n.

Voor AFA en GFA, waar we met maximum likelihoodschatters werken, is het uitgangspunt de volgende stelling: (SILVEY (1970)):

Stelling 3.2.1.

Zij $\theta = (\theta_1,\ldots,\theta_k)'$ een k-dimensionale parameter van een dichtheid f, die de kansverdeling van een p-dimensionale vector $\underline{x}$ beschrijft.
Zij $\underline{\hat{\theta}}_n$ de maximum likelihoodschatter van $\theta$, onder de r-voudige bijvoorwaarde $g(\theta) \overset{\text{def}}{=} (g_1(\theta),g_2(\theta),\ldots,g_r(\theta))' = 0$, op grond van n trekkingen uit f.
Zij

$$J(\theta) = \left[J(\theta_i,\theta_j)\right]_{i,j = 1,\ldots,k} \overset{\text{def}}{=} \left[-E\left(\frac{\partial^2 \log f(x)}{\partial\theta_i \partial\theta_j}\right)\right] i,j = 1,\ldots,k$$

(Fisher informatiematrix)

Zij

$$\tilde{J}(\theta) = \begin{bmatrix} J(\theta) & \frac{\partial g(\theta)}{\partial \theta} \\ \left(\frac{\partial g(\theta)}{\partial \theta}\right)' & 0 \end{bmatrix}$$

Zij $A(\theta)$ het $(k \times k)$-linkerbovenvierkant van $(\tilde{J}(\theta))^{-1}$ dan geldt

$$\operatorname{cov}(\hat{\underline{\theta}}_n) \sim \frac{1}{n} A(\theta), \quad n \to \infty$$

(onder passende regulariteitsvoorwaarden voor f en g).

Deze stelling is van nut wanneer f de dichtheid van de multivariaat normale verdeling is met covariantiematrix $\Sigma$, omdat de uitdrukking voor de informatiematrix dan tot hanteerbare uitdrukkingen leidt: voor elk tweetal parameters $\alpha$ en $\beta$ geldt

$$J(\alpha,\beta) \overset{\text{def}}{=} -E\frac{\partial^2 \log f(\underline{x})}{\partial\alpha\partial\beta} = \tfrac{1}{2}\text{spoor}(\Sigma^{-1}\frac{\partial\Sigma}{\partial\alpha}\Sigma^{-1}\frac{\partial\Sigma}{\partial\beta})$$

Wanneer we dit voor AFA gaan toepassen is $\theta$ een $(pm + p)$-dimensionale parameter, bestaande uit de $\lambda_{ij}(i = 1,\ldots p; j = 1,\ldots m)$ en de $\psi_i(i = 1,\ldots,p)$ in een zekere volgorde (bijv. $\lambda_{ir}(1 \le i \le p,\ 1 \le r \le m)$, $\psi_i(1 \le i \le p)$) terwijl de bijvoorwaarde afhangt van de gekozen rotatie (zie 2.13).
(Men kan ook zogenaamde "scheve rotatie" met deze techniek behandelen, hoewel volgens JENNRICH (1974) de formules dan zeer ingewikkeld worden. Wij gaan hier niet op in.)

We onderscheiden:

*AFA voor een covariantiematrix*
dan geldt

$$\left.\begin{aligned} J(\lambda_{ir},\lambda_{js}) &= \sigma^{ij}(\Lambda'\Sigma^{-1}\Lambda)_{rs} + (\Sigma^{-1}\Lambda)_{is}(\Sigma^{-1}\Lambda)_{jr} \\ J(\lambda_{ir},\psi_j) &= \sigma^{ij}(\Sigma^{-1}\Lambda)_{jr} \\ J(\psi_i,\psi_j) &= \tfrac{1}{2}(\sigma^{ij})^2 \end{aligned}\right\} \begin{matrix} 1 \le i,j \le p \\ 1 \le r,s \le m \end{matrix}$$

waarin $\sigma^{ij} \overset{\text{def}}{=} (\Sigma^{-1})_{ij}$.

De bijvoorwaarde hangt af van de rotatie:

*voor de canonieke oplossing:*

De canonieke oplossing beantwoordt aan: $\Lambda'\Psi^{-1}\Lambda$ is diagonaal. Met de notatie $g_{uv} \overset{\text{def}}{=} (\Lambda'\Psi^{-1}\Lambda)_{uv}$ $(1 \le u \le v \le m)$ komt dat neer op $g_{uv} = 0$. Zij $\delta_{ij}$ de Kroneckerdelta, dan geldt:

$$\left.\begin{aligned} \frac{\partial g_{uv}}{\partial \lambda_{ir}} &= (\delta_{ru}\lambda_{iv} + \delta_{rv}\lambda_{iu})\psi_i^{-1} \\ \frac{\partial g_{uv}}{\partial \psi_i} &= -\lambda_{iu}\lambda_{iv}\psi_i^{-2} \end{aligned}\right\} \begin{matrix} 1 \le u \le v \le m \\ 1 \le i \le p \\ 1 \le r \le m \end{matrix}$$

*voor de varimax, quartimax en equimax oplossingen (niet genormaliseerd):*

Voer in

$$a_{iuv} \stackrel{\text{def}}{=} \lambda_{iv}^3 - 3\lambda_{iu}^2\lambda_{iv} - \frac{c}{p}\left[\lambda_{iv}\{(\Lambda'\Lambda)_{vv} - (\Lambda'\Lambda)_{uu}\} - 2\lambda_{iu}(\Lambda'\Lambda)_{uv}\right]$$

met $c = 0$ voor quartimax

$c = 1$ voor varimax

$c = \frac{m}{2}$ voor equimax

en $$\left.\begin{aligned} \frac{\partial g_{uv}}{\partial \lambda_{ir}} &= \delta_{ur}\, a_{iuv} - \delta_{vr}\, a_{ivu} \\ \frac{\partial g_{uv}}{\partial \psi_i} &= 0 \end{aligned}\right\} \begin{matrix} 1 \le i \le p \\ 1 \le r \le m \\ 1 \le u \le v \le m \end{matrix}$$

*voor de genormaliseerde varimax, quartimax en equimax oplossingen*

Hierbij worden de langs analoge lijnen te verkrijgen formules zo monstrueus dat we ze hier weglaten.

*AFA voor de correlatiematrix*

Hiertoe beschouwen we de correlatiematrix $\Gamma$, die ontbonden wordt als $\Lambda\Lambda' + (I - \text{diag}\Lambda\Lambda')$ als geparametriseerd door $\Lambda$ en door de wortels uit de diagonaal elementen van $\Sigma$, de covariantiematrix, genoteerd als $(\sigma_1,\ldots,\sigma_p)'$. De elementen van de informatiematrix zijn:

$$J(\lambda_{ir},\lambda_{js}) \stackrel{\text{def}}{=} \gamma^{ij}(\Lambda'\Gamma^{-1}\Lambda)_{rs} + (\Gamma^{-1}\Lambda)_{is}(\Gamma^{-1}\Lambda)_{jr} - 2\gamma^{ij}\lambda_{ir}(\Gamma^{-1}\Lambda)_{is} +$$

$$-2\gamma^{ij}\lambda_{js}(\Gamma^{-1}\Lambda)_{jr} + 2(\gamma^{ij})^2\lambda_{ir}\lambda_{js}$$

$$J(\lambda_{ir},\sigma_j) \stackrel{\text{def}}{=} \gamma^{ij}\lambda_{jr} + \delta_{ij}(\Gamma^{-1}\Lambda)_{jr} - 2\delta_{ij}\gamma^{ii}\lambda_{ir}$$

$$J(\sigma_i,\sigma_j) \overset{\text{def}}{=} \delta_{ij} + \gamma_{ij}\gamma^{ij}$$

voor $1 \le i,j \le p$, $1 \le r,s \le m$ en met $\gamma^{ij} = (\Gamma^{-1})_{ij}$ en $\delta_{ij}$ de Kronecker-delta.

De bijvoorwaarde leidt dan tot

*voor de canonieke rotatie*

$$\left.\begin{aligned} \frac{\partial g_{uv}}{\partial \lambda_{ir}} &= \frac{\delta_{ur}\lambda_{iv} + \delta_{vr}\lambda_{iu}}{1 - \sum_{j=1}^{m} \lambda_{ij}^2} + \frac{2\lambda_{iu}\lambda_{iv}\lambda_{ir}}{(1 - \sum_{j=1}^{m} \lambda_{ij}^2)^2} \\ \frac{\partial g_{uv}}{\partial \sigma_j} &= 0 \end{aligned}\right\} \begin{aligned} &1 \le u \le v \le m \\ &1 \le i \le p \\ &1 \le r \le m \end{aligned}$$

*voor de varimax, quartimax en equimax oplossingen*

Voor deze oplossingen zijn de afgeleiden van de uitdrukkingen die in de bijvoorwaarden op nul gesteld worden gelijk aan die welke bij het geval van de covariantiematrix optreden.

*GFA voor een covariantiematrix*

De elementen van de informatiematrix zijn, als $\Sigma = \Lambda\Lambda' + \psi I$, met $\psi > 0$

$$\left.\begin{aligned} J(\lambda_{ir},\lambda_{js}) &= \sigma^{ij}(\Lambda'\Sigma^{-1}\Lambda)_{rs} + (\Sigma^{-1}\Lambda)_{is}(\Sigma^{-1}\Lambda)_{jr} \\ J(\lambda_{ir},\psi) &= \sum_{j=1}^{p} \sigma^{ij}(\Sigma^{-1}\Lambda)_{jr} \\ J(\psi,\psi) &= \tfrac{1}{2}\text{spoor}(\Sigma^{-2}). \end{aligned}\right\} \begin{aligned} &1 \le i,j \le p \\ &1 \le r,s \le m \end{aligned}$$

De bijvoorwaarde leidt tot

*in het canonieke geval*

$$\left.\begin{aligned} \frac{\partial g_{uv}}{\partial \lambda_{ir}} &= \delta_{ru}\lambda_{iv} + \delta_{rv}\lambda_{iu} \\ \frac{\partial g_{uv}}{\partial \psi} &= 0 \end{aligned}\right\} \begin{aligned} &1 \le u \le v \le m \\ &1 \le r,s \le m \end{aligned}$$

*Voor de varimax, quartimax en equimax oplossingen*

Voor deze oplossingen zijn de afleidingen verkregen van de bijvoorwaarden gelijk aan die voor het AFA-geval.

*GFA voor een correlatiematrix*

Dit geval is door ons niet uitgewerkt.

*Jøreskog's model*

Daar in JFA geen gebruik gemaakt wordt van maximum likelihoodschatters, is bovenstaande benadering niet toepasbaar. JØRESKOG (1963) leidt de asymtotische (co)varianties van de gebruikte schatters af en geeft een approximatie die wat beter hanteerbaar is. Deze laatste vermelden wij hier, *voor de canonieke oplossing* bij het geval van een *covariantiematrix*.

Notatie: $\Delta \overset{\text{def}}{=} \text{diag}\Sigma^{-1}$;

$\Sigma^* \overset{\text{def}}{=} \Delta^{\frac{1}{2}}\Sigma\Delta^{\frac{1}{2}}$;

$\delta_1,\ldots,\delta_p$ de eigenwaarden in aflopende volgorde van $\Sigma^*$;

$\theta$ de gemeenschappelijke waarde van $\delta_{m+1},\ldots,\delta_p$;

$\lambda_i$ de i-de kolom uit $\Lambda$, de canonieke oplossing (i.e. $\Lambda'\Delta\Lambda$ diag);

$\underline{\ell}_i$ de i-de kolom uit $\underline{L}$ (vgl. 3.1), de canonieke oplossing.

Nu geldt:

$$E(\underline{\ell}_i - \lambda_i)(\underline{\ell}_i - \lambda_i)' \approx$$

$$\frac{1}{n}\ \frac{\delta_i}{\delta_i - \theta}\{\Sigma - \frac{\delta_i}{2(\delta_i - \delta)}\lambda_i\lambda_i' + \sum_{j\neq i}\frac{\delta_j}{(\delta_j - \theta)}\cdot\left[\left(\frac{\delta_j - \theta}{\delta_i - \delta_j}\right)^2 - 1\right]\lambda_j\lambda_j'\};$$

$$i = 1,\ldots,m$$

en

$$E(\underline{\ell}_i - \lambda_i)(\underline{\ell}_j - \lambda_j)' \approx -\frac{1}{n}\frac{\delta_i\delta_j}{(\delta_i - \delta_j)^2}\lambda_j\lambda_i';\ i,j = 1,\ldots,m;\ i \neq j$$

Voor gebruik in de praktijk moet men voor $\Sigma$ $\underline{S}$, voor $\lambda_i$ $\underline{\ell}_i$, voor $\delta_i$ $\underline{c}_i$, voor $\theta$ $\underline{t}$ invullen, zoals gedefinieerd in 3.1 bij JFA. Asymptotische (co)varianties van de schatters van de ontbinding van de correlatiematrix en van de factorladingen bij andere rotaties dan de canonieke, zijn waarschijnlijk veel lastiger te berekenen en dit is ook niet door Jøreskog gedaan.

## 3.3. GEDRAG VAN EIGENWAARDEN EN EIGENVECTOREN BIJ SCHAALTRANSFORMATIES

Zij $\Sigma$ de covariantiematrix van een p-variate stochastische variabele $\underline{x}$, en zij D een $(p \times p)$-diagonaalmatrix met positieve diagonaalelementen, dan is $\Sigma^* \overset{def}{=} D\,\Sigma\,D$ de covariantiematrix van de stochastische variabele $D\underline{x}$, d.w.z. de stochastische variabele $\underline{x}$ nadat een schaaltransformatie is uitgevoerd, dus $\underline{x}_i$ gaat over in $d_{ii}\underline{x}_i$ $(i = 1,\ldots p)$. We kunnen ons afvragen welke de veranderingen van de eigenwaarden en eigenvectoren van $\Sigma^*$ t.o.v. die van $\Sigma$ zijn bij verschillende keuze van D.

Deze vraag is van belang voor principale componentenanalyse omdat de componenten direct bepaald worden door de eigenvectoren. Daar is deze vraag dus equivalent met de vraag: hoe gevoelig is principale componentenanalyse voor schaalverandering? (vgl. 2.9). Om technische redenen is de vraag ook interessant bij maximum likelihoodoplossing van AFA (vgl. 3.7).

Helaas is het een zeer moeilijke vraag. Voor $p = 2$ kan men rechtstreeks uitrekenen dat elke vector als eigenvector kan optreden door geschikte keuze van D. Voor $p \geq 3$ is, voor zover ons bekend, de vraag onopgelost. Wij demonstreren hier een bijzonder geval, waar sommige schaalfactoren, zeg de eerste r, naar oneindig gaan en de overige naar constanten convergeren. Zonder verlies van algemeenheid laten wij deze laatste $p - r$ diagonaalelementen constant gelijk aan 1 blijven.

Splits alle betrokken matrices op in submatrices van omvang $r \times r$, $r \times (p-r)$, $(p-r) \times r$ en $(p-r) \times (p-r)$:

$$\Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right], \quad D = \left[\begin{array}{c|c} D_1 & 0 \\ \hline 0 & I \end{array}\right], \quad \underline{x} = \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_p \end{bmatrix}$$

$$\Sigma^* \overset{def}{=} D\,\Sigma\,D = \left[\begin{array}{c|c} D_1\Sigma_{11}D_1 & D_1\Sigma_{12} \\ \hline \Sigma_{21}D_1 & \Sigma_{22} \end{array}\right]$$

Wij bekijken thans het asymptotisch gedrag van de eigenwaarden en eigenvectoren van $\Sigma^*$ als $d_{ii} \to \infty$, $i = 1,\ldots,r$.

Gebruikmakend van de notatie voor de matrix van partiële covarianties van $\underline{x}_i$ gegeven $\underline{x}_j$:

$$\Sigma_{11\cdot 2} \overset{\text{def}}{=} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$
$$\Sigma_{22\cdot 1} \overset{\text{def}}{=} \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

en een bekende stelling over de inverse van gepartitioneerde matrices (zie eventueel MORRISON (1976)) volgt:

$$(\Sigma^*)^{-1} = \left[\begin{array}{c|c} D_1^{-1}\Sigma_{11\cdot 2}^{-1}D_1^{-1} & -D_1^{-1}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22\cdot 1}^{-1} \\ \hline -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11\cdot 2}^{-1}D_1^{-1} & \Sigma_{22\cdot 1} \end{array}\right]$$

Hieruit zien wij dat

$$(\Sigma^*)^{-1} \to \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & \Sigma_{22\cdot 1} \end{array}\right] \quad \text{als } d_{ii} \to \infty,\ 1 \le i \le r.$$

Merk nu op (zie eventueel WILKINSON (1965))

a. eigenvectoren van $(\Sigma^*)^{-1}$ zijn eigenvectoren van $\Sigma^*$;
b. eigenwaarden van $(\Sigma^*)^{-1}$ zijn inverses van eigenwaarden van $\Sigma^*$;
c. eigenwaarden zijn continue functies van de elementen van een matrix;
d. niet multipele eigenvectoren zijn continue functies van de elementen van een matrix.

Laten de eigenwaarden van $\Sigma_{22\cdot 1}$ in aflopende volgorde $\theta_{r+1},\ldots,\theta_p$ zijn, met bijbehorende eigenvectoren $v_{r+1},\ldots,v_p$ (dit zijn vectoren van $p - r$ elementen).
Laat de eigenwaarden van $\Sigma^*$ in aflopende volgorde $\zeta_1,\ldots,\zeta_p$ zijn, met eigenvectoren $u_1,\ldots,u_p$.
Dan volgt uit het bovenstaande

$$\left.\begin{array}{l} \zeta_i \to \infty,\ 1 \le i \le r \\ \zeta_i \to \dfrac{1}{\theta_i},\ r < i \le p \\ u_i \to \begin{pmatrix} 0 \\ \vdots \\ v_i \end{pmatrix},\ r < i \le p \\ \text{de ruimte opgespannen door } u_1,\ldots,u_r \to \\ \text{de ruimte opgespannen door de eerste} \\ r \text{ kolommen van de } (p \times p)\text{-eenheidsmatrix} \end{array}\right\} \text{ als } d_{ii} \to \infty;\quad 1 \le i \le r$$

Voor principale componentenanalyse betekent dit dat men de eerste r componenten tezamen willekeurig veel op de eerste r variabelen kan laten lijken door deze variabelen een schaaltransformatie te laten ondergaan (zie 2.9).

## 3.4. DE VOORBEELDEN VAN WILSON & WORCESTER

Zoals in 2.4 uiteengezet hebben WILSON & WORCESTER (1939) voorbeelden gegeven van een drietal correlatiematrices $\Gamma_1,\Gamma_2,\Gamma_3$ die sprekend op elkaar lijken en waarvan de eerste twee, de tweede één en de derde geen exacte 3-factoroplossingen voor het AFA-model (met verschillende $\Psi$) toelaat. Omdat het geciteerde artikel moeilijk toegankelijk is, geven wij deze voorbeelden hier weer. Terwijl Wilson & Worcester als oplossing telkenmale die $\Lambda$ kiezen uit alle rotaties die nullen boven de diagonaal heeft (Thurstone's methode), geven wij als oplossing de tegenwoordig meer populaire genormaliseerde varimaxrotatie (zie 2.13).

Bovendien laten we zien dat een zwak oplossingsmechanisme, zoals de indirecte principale componentenmethode (zie 2.3) de correcte oplossing(en) niet vindt, alhoewel hier dus exacte oplossingen bestaan. Hierbij is gebruik gemaakt van het programmapakket SPSS (NIE e.a. (1975) en wel de factormethode "PA2", waarbij we het juiste aantal factoren opgaven, omdat er anders helemaal niets van terecht kwam.

$$\Gamma_1 = \begin{bmatrix} 1 & & & & \\ 0,56 & 1 & & & \\ 0,16 & 0,20 & 1 & & \\ 0,48 & 0,66 & 0,18 & 1 & \\ 0,24 & 0,51 & 0,07 & 0,30 & 1 \\ 0,64 & 0,86 & 0,23 & 0,72 & 0,41 \quad 1 \end{bmatrix} \qquad \Gamma_2 = \begin{bmatrix} 1 & & & & \\ 0,56 & 1 & & & \\ 0,16 & 0,20 & 1 & & \\ 0,48 & 0,66 & \underline{0,08} & 1 & \\ 0,24 & 0,51 & \underline{0,27} & 0,30 & 1 \\ 0,64 & 0,86 & \underline{0,13} & 0,72 & 0,41 \quad 1 \end{bmatrix}$$

$$\Gamma_3 = \begin{bmatrix} 1 & & & & \\ 0,56 & 1 & & & \\ 0,16 & 0,20 & 1 & & \\ 0,48 & 0,66 & 0,18 & 1 & \\ 0,24 & 0,51 & 0,07 & 0,30 & 1 \\ 0,64 & 0,86 & \underline{0,25} & \underline{0,76} & 0,33 \quad 1 \end{bmatrix}$$

In $\Gamma_2$ en $\Gamma_3$ zijn onderstreept de elementen die van de overeenkomstige elementen in $\Gamma_1$ verschillen.

tabel 3.4.1

Drie correlatiematrices

| oplossing 1 | | | | | oplossing 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| variabele | communaliteit | factor I | II | III | variabele | communaliteit | factor I | II | III |
| 1 | 0,64 | 0,41 | 0,16 | 0,67 | 1 | 0,43 | 0,61 | 0,24 | 0,05 |
| 2 | 0,85 | 0,69 | 0,55 | 0,29 | 2 | 0,90 | 0,67 | 0,68 | -0,06 |
| 3 | 0,06 | 0,22 | 0,04 | 0,10 | 3 | 0,06 | 0,25 | 0,06 | -0,02 |
| 4 | 0,56 | 0,66 | 0,25 | 0,26 | 4 | 0,55 | 0,67 | 0,32 | 0,01 |
| 5 | 0,50 | 0,16 | 0,68 | 0,10 | 5 | 0,39 | 0,16 | 0,60 | 0,03 |
| 6 | 0,93 | 0,81 | 0,35 | 0,37 | 6 | 1,00 | 0,87 | 0,44 | 0,23 |

door SPSS gevonden oplossing *)

| variabele | communaliteit | factor I | II | III |
|---|---|---|---|---|
| 1 | 0,47 | 0,62 | 0,24 | 0,16 |
| 2 | 0,89 | 0,64 | 0,69 | -0,01 |
| 3 | 0,06 | 0,24 | 0,06 | -0,01 |
| 4 | 0,57 | 0,67 | 0,33 | -0,09 |
| 5 | 0,37 | 0,16 | 0,59 | 0,01 |
| 6 | 0,93 | 0,84 | 0,46 | 0,04 |

tabel 3.4.2

De varimax geroteerde factoroplossingen van $\Gamma_1$

| oplossing | | | | | door SPSS gevonden oplossing **) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| variabele | communaliteit | factor I | II | III | variabele | communaliteit | factor I | II | III |
| 1 | 0,64 | 0,47 | 0,17 | 0,62 | 1 | 0,54 | 0,69 | -0,03 | 0,24 |
| 2 | 0,85 | 0,83 | 0,37 | 0,17 | 2 | 0,86 | 0,76 | 0,48 | 0,21 |
| 3 | 0,21 | 0,02 | 0,44 | 0,12 | 3 | 0,27 | 0,06 | 0,10 | 0,51 |
| 4 | 0,56 | 0,71 | 0,09 | 0,20 | 4 | 0,56 | 0,70 | 0,25 | 0,02 |
| 5 | 0,50 | 0,35 | 0,61 | -0,05 | 5 | 0,48 | 0,23 | 0,51 | 0,41 |
| 6 | 0,93 | 0,90 | 0,17 | 0,30 | 6 | 0,93 | 0,91 | 0,32 | 0,09 |

tabel 3.4.3

De varimax geroteerde factoroplossingen van $\Gamma_2$

Geval $\Gamma_3$

De methode "PA2" van SPSS weet in dit geval inderdaad geen driefactoroplossing te produceren: na 6 iteraties is een van de specifiteiten kleiner dan nul geschat; het programma is daartegen beveiligd en stopt. SPSS suggereerde zelf een 1-factoroplossing met maximumresidu 0,125.

## 3.5. BEREKENING VAN DE GUTTMAN CRITERIUMWAARDE EN CONSTRUCTIE VAN MAXIMAAL VERSCHILLENDE FACTOREN

Deze paragraaf is een technische uitwerking van 2.12. De formule van de Guttman criteriumwaarde (GCW) wordt afgeleid en een constructie van stellen factoren die onderling maximaal verschillen wordt gegeven.

---

*) Wanneer niet expliciet om drie factoren gevraagd wordt, geeft de SPSS-methode PA2 er één. De hier gegeven oplossing vereiste 20 iteratieslagen en de maximumafwijking van de laatste iteratieslag t.o.v. de voorlaatste is $_{10}{}^{-3}$ in de factorladingen. De reproduktie van $\Gamma_1$ door deze "oplossing" is zeer goed te noemen, maximum residu < 0,005 (d.w.z. max $|\Gamma_{1ij} - (\Lambda\Lambda' + \Psi)_{ij}| < 0{,}005$).

**) Indien SPSS het voor het zeggen heeft krijgt men een tweefactoroplossing na niet minder dan 53 iteraties. De driefactoroplossing vereist 15 iteraties en geeft een maximum van < 0,002. □

Zij zonder verlies van algemeenheid $\underline{x} \overset{\text{def}}{=} (\underline{x}_1,\ldots,\underline{x}_p)'$ zodanig dat $E\underline{x} = 0$ en $\Sigma \overset{\text{def}}{=} E(\underline{x}\underline{x}') = \Lambda\Lambda' + \Psi$ voor zekere $(p \times m)$-matrix $\Lambda$ en een $(p \times p)$-diagonaal matrix $\Psi$ met positieve diagonaalelementen.

LEMMA 3.5.1. *Als er een* m*-dimensionale* $\underline{f}$ *en een* p*-dimensionale* $\underline{s}$ *bestaan met:*

$$
\begin{aligned}
&E\underline{f} = 0,\ E\underline{s} = 0,\ E\underline{s}\underline{f}' = 0 \\
&E\underline{f}\underline{f}' = I,\ E\underline{s}\underline{s}' = \Psi \\
&\qquad \text{en } \underline{x} = \underline{f} + \underline{s}
\end{aligned} \tag{1}
$$

*dan bestaat er een* m*-dimensionale stochastische variabele* $\underline{y}$ *met de eigenschappen:*

$$
\begin{aligned}
&E\underline{y} = 0,\ E\underline{y}\underline{x}' = 0 \\
&E\underline{y}\underline{y}' = I - \Lambda'\Sigma^{-1}\Lambda
\end{aligned} \tag{2}
$$

*èn*

$$
\begin{aligned}
&\underline{f} = \Lambda'\Sigma^{-1}\,\underline{x} + y \\
&\underline{s} = \underline{x} - \Lambda\underline{f}
\end{aligned} \tag{3}
$$

BEWIJS. Voor iedere $(p \times 1)$-vector $a \neq 0$ geldt

$$
\begin{aligned}
a'\Sigma a &= a'\Lambda\Lambda'a + a'\Psi a \\
&= (\Lambda'a)'(\Lambda'a) + a'\Psi a \\
&= \sum_j [(\Lambda'a)_j]^2 + \sum_i a_i^2\,\psi_i > 0, \text{ daar alle } \psi_i > 0.
\end{aligned}
$$

Dus $\Sigma$ is positief definiet, waaruit volgt, dat $\Sigma^{-1}$ bestaat. Dan is het mogelijk $\underline{y}$ en $\underline{s}$ d.m.v. (3) te definiëren en (2) volgt direct door uitschrijven. □

Omgekeerd geldt ook:

LEMMA 3.5.2. *Wanneer* $\underline{y}$ *een* m*-dimensionale stochastische variabele is, die aan* (2) *voldoet en* $\underline{f}$ *en* $\underline{s}$ *door* (3) *gedefinieerd worden, dan voldoen* $\underline{f}$ *en* $\underline{s}$ *aan* (1).

BEWIJS. Dit volgt direct door uitschrijven. □

Tenslotte garandeert het volgende lemma het bestaan van $\underline{y}$ zoals in lemma 3.5.2 bedoeld:

LEMMA 3.5.3. *Er bestaan* m*-dimensionale stochastische variabelen* $\underline{y}$*, die aan* (2) *voldoen.*

BEWIJS. $I + \Lambda'\Psi^{-1}\Lambda$ is positief definiet (bewijs analoog als in 3.5.1). $\Sigma^{-1}$ bestaat (bewijs gelijk aan dat in 3.5.1). Merk op dat $(I + \Lambda'\Psi^{-1}\Lambda)(I - \Lambda'\Sigma^{-1}\Lambda) = I$, dus $(I - \Lambda'\Sigma^{-1}\Lambda) = (I + \Lambda'\Psi^{-1}\Lambda)^{-1}$ en derhalve ook positief definiet. Dan bestaat T z.d.d. $TT' = I - \Lambda'\Sigma^{-1}\Lambda$. Zij nu $\underline{e}$ een m-dimensionale stochastische variabele met $E\underline{e} = 0$ en $E\underline{e}\underline{e}' = I$, die ongecorreleerd is met $\underline{x}$, d.w.z. $E\underline{e}\underline{x}' = 0$; dan voldoet $\underline{y} \overset{\text{def}}{=} T\underline{e}$ aan (2). □

GEVOLG. Daar men willekeurig veel verschillende $\underline{e}$ kan maken (met een aselector) kan men zo willekeurig veel verschillende $\underline{y}$ en daarmee $\underline{f}$ en $\underline{s}$ construeren.

Deze lemma's gebruiken we nu om bij een gegeven stel factoren een ander stel te construeren, dat daar minimaal mee correleert. Zij $\underline{f}'$ zo'n rijtje factoren, definieer $\underline{y} = \underline{f} - \Lambda'\Sigma^{-1}\underline{x}$ en definieer $\underline{y}^* = -\underline{y}$; dan voldoet $\underline{y}^*$ aan (2) en dus is $\underline{f}^* \overset{\text{def}}{=} \Lambda'\Sigma^{-1}\underline{x} + \underline{y}^*$ ook een stelsel factoren, terwijl onder alle variabelen van de vorm $(\Lambda'\Sigma^{-1}\underline{x})_j + \underline{z}_j$, $\underline{z}_j$ ongecorreleerd met $\underline{x}$ en met de gevraagde variantie, juist $(\Lambda'\Sigma^{-1}\underline{x})_j + (-\underline{y}_j)$ het laagst gecorreleerd is met $(\Lambda'\Sigma^{-1}\underline{x})_j + \underline{y}_j$; $j = 1,2,\ldots,m$.
De Guttman criteriumwaarden (GCW's) zijn juist deze correlatiecoëfficienten, dus de diagonaalelementen van $E(\underline{f}^*\underline{f}')$.

$$\underline{f}^* = \Lambda'\Sigma^{-1}\underline{x} - \underline{y} = \Lambda'\Sigma^{-1}\underline{x} - (\underline{f} - \Lambda'\Sigma^{-1}\underline{x}) = 2\Lambda'\Sigma^{-1}\underline{x} - \underline{f}$$

dus

$$E(\underline{f}^*\underline{f}') = E((2\Lambda'\Sigma^{-1}\underline{x} - \underline{f})\underline{f}') = 2\Lambda'\Sigma^{-1}\Lambda - I.$$

We noemen deze matrix $G(\Lambda)$; de GCW voor de j-de factor corresponderend met

de factorladingenmatrix $\Lambda$ is $G(\Lambda)_{jj}$.

Merk op dat dit voor iedere $\Lambda$ die aan $\Sigma = \Lambda\Lambda' + \Psi$ voldoet geldt, i.h.b. ook voor geroteerde oplossingen, dus

$$G(\Lambda\Theta) = 2(\Lambda\Theta)'\Sigma^{-1}(\Lambda\Theta) - I = \Theta'(2\Lambda'\Sigma^{-1}\Lambda - I)\,\Theta = \Theta'G(\Lambda)\Theta.$$

Dit kan men soms gebruiken om de GCW's voor een geroteerde oplossing uit die voor de originele oplossing te berekenen.

*Berekening van G*

De uitdrukking $G(\Lambda) = 2\Lambda'\Sigma^{-1}\Lambda - I$ is meestal niet de eenvoudigste manier om $G(\Lambda)$ te berekenen. Wanneer $\Lambda$ de canonieke oplossing is, zijn de volgende formules vaak handig.

$$G(\Lambda) = 2\Lambda'\Sigma^{-1}\Lambda - I = I - 2(I - \Lambda'\Sigma^{-1}\Lambda) = I - 2(I + \Lambda'\Psi^{-1}\Lambda)^{-1}$$

en dit is een diagonaalmatrix voor de canonieke oplossing ($\Lambda'\Psi^{-1}\Lambda$ diagonaal) terwijl bovendien geldt dat de diagonaalelementen van $\Lambda'\Psi^{-1}\Lambda$ 1 kleiner zijn dan de m grootste eigenwaarden van $\Psi^{-\frac{1}{2}}\Sigma\Psi^{-\frac{1}{2}}$ die in AFA in de berekening van $\Lambda$ en $\psi$ voorkomen.
Noemen we deze eigenwaarden $\theta_1,\ldots,\theta_m$, dan geldt

$$[G(\Lambda)]_{jj} = 1 - 2\,\frac{1}{1 + \theta_j - 1} = 1 - \frac{2}{\theta_j} \quad \text{(canonieke } \Lambda\text{, AFA)}.$$

Bij berekening van het GFA-model worden meestal de eigenwaarden $d_1,\ldots,d_m$ van $\Sigma$ i.p.v. $\theta_1\ldots\theta_m$ van $\Psi^{-\frac{1}{2}}\Sigma\Psi^{-\frac{1}{2}}$ opgegeven. De relatie $\theta_j = \frac{dj}{\sigma^2}$ helpt ons verder

$$[G(\Lambda)]_{jj} = 1 - 2\,\frac{\sigma^2}{d_j} \quad \text{(canonieke } \Lambda\text{, GFA)}.$$

In JFA worden meestal de eigenwaarden $c_1,\ldots,c_m$ van $(\text{diag}\ \Sigma^{-1})^{\frac{1}{2}}\Sigma(\text{diag}\ \Sigma^{-1})^{\frac{1}{2}}$ opgegeven i.p.v. $\theta_1,\ldots,\theta_m$, maar de relatie $\theta_j = \frac{c_j}{\theta}$ helpt ons verder:

$$[G(\Lambda)]_{jj} = 1 - 2\,\frac{\theta}{c_j} \quad \text{(canonieke } \Lambda\text{, JFA)}.$$

Tenslotte moeten we hierin schattingen substitueren. Met de notatie uit 3.1 krijgen we dan, als $\underline{g}_j$ de schatter voor de GCW voor de j-de canonieke factor is:

$$\left.\begin{aligned} &\text{AFA: } \underline{g}_j = 1 - \frac{2}{\underline{\theta}_j}, \text{ met } \underline{\theta}_j \text{ de j-de eigenwaarde van } \underline{P}^{-\frac{1}{2}}\underline{S}\underline{P}^{-\frac{1}{2}} \\ &\text{GFA: } \underline{g}_j = 1 - \frac{2\underline{s}^2}{\underline{d}_j} \\ &\text{JFA: } \underline{g}_j = 1 - \frac{2\underline{t}}{\underline{c}_j} \end{aligned}\right\} j = 1,\ldots,m$$

Tenslotte volgt, dat na rotatie met rotatiematrix $\Theta$ (i.e. $\Lambda = \Lambda_{\text{canoniek}}\Theta$ en $\Theta\Theta' = I$) de Guttman criteriumwaarden $\underline{h}_1,\ldots,\underline{h}_m$ van de geroteerde oplossing gegeven worden door

$$\begin{bmatrix} \underline{h}_1 & & & 0 \\ & \underline{h}_2 & & \\ & & \ddots & \\ 0 & & & \underline{h}_m \end{bmatrix} = \Theta' \begin{bmatrix} \underline{g}_1 & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & \underline{g}_m \end{bmatrix} \Theta$$

Bij gecorreleerde factoren zijn overeenkomstige formules te verkrijgen. Dus als $\underline{x} = \Lambda\underline{f} + \underline{s}$ en $E(\underline{f}\underline{f}') = \Phi$, waarbij $\Phi$ positief definiet is en $\text{diag}(\Phi) = I$, is de GCW van de j-de factor gelijk aan het j-de diagonaalelement van $G(\Lambda) = 2\Phi\Lambda'\Sigma^{-1}\Lambda\Phi - \Phi$. Bij een transformatie

$$\underline{f} \to \underline{f}^* = A^{-1}\underline{f}$$

zodanig dat $\text{diag } A^{-1}\Phi A^{-1'} = I \qquad (\text{cov}(\underline{f}^*) = A^{-1}\Phi A^{-1'})$

zodat $\qquad \Lambda \to \Lambda^* = \Lambda A$ en $\underline{x} = \Lambda^*\underline{f}^* + \underline{s}$;

geldt bovendien $G(\Lambda A) = A^{-1}G(\Lambda)A^{-1'}$.

*Verband Guttman criterium en factorscores*

In formule (3) van deze paragraaf stond

$$\underline{f} = \Lambda'\Sigma^{-1}\underline{x} + \underline{y}$$

waarbij $\underline{y}$ ongecorreleerd is met $\underline{x}$.
Merk verder op, dat $G(\Lambda) = I - 2\text{cov}(\underline{y})$.

Factorscores zijn schattingen van de waarde aangenomen door $\underline{f}$, gegeven dat $\underline{x} = x$. Als $\sum$ bekend is, is één bekende methode ("regressiemethode")

$$\hat{\underline{f}} = \Lambda'\Sigma^{-1}\underline{x}; \text{ oftewel } \hat{\underline{f}} = \underline{f} - \underline{y}.$$

Andere methoden gebruiken $\hat{\hat{\underline{f}}} = A\hat{\underline{f}}$ voor een bepaalde matrix A, gekozen aan de hand van een gegeven optimaliteitscriterium en/of een gegeven voorwaarde (bijv. $E(\hat{\hat{f}}\hat{\hat{f}}') = I$).

Het is nu intuïtief duidelijk, dat, naarmate de varianties van de componenten van $\underline{y}$ groter zijn, de schatting van $\underline{f}$, gegeven $\underline{x}$, moeilijker is en de GCW's kleiner zijn.

## 3.6. ONEIGENLIJKE OPLOSSINGEN

Men spreekt van een oneigenlijke factoroplossing (Heywood case in de Engelstalige literatuur) wanneer schattingen van sommige uniciteiten, d.w.z. varianties van specifieke gedeelten, kleiner dan of gelijk aan nul uitvallen, of naar (of voorbij) nul dreigen te convergeren in de loop van een iteratieve schattingsprocedure. Als men gebruik maakt van een iteratieve methode, komen oneigenlijke oplossingen in de praktijk veel voor, iets dat niet zonder meer als slechte eigenschap van een dergelijke methode is te beschouwen.

Dergelijke oneigenlijke oplossingen leveren vele problemen, zoals

a. Verschillende oplossingsalgoritmen krijgen numerieke problemen, als een oneigenlijke oplossing benaderd wordt. Doorgaans wordt dit veroorzaakt doordat $\Psi^{-1}$ moet worden berekend, waarvan bepaalde elementen naar $\infty$ divergeren.
b. Andere algorithmen gaan door en leveren als schattingen negatieve uniciteiten, d.w.z. negatieve varianties. Deze zijn niet te interpreteren.
c. Als elementen van $\Psi$ *in werkelijkheid* nul zijn, gelden de gebruikelijke

formules voor asymptotische verdeling van de maximum likelihoodschatters niet, evenmin als de asymptotische verdeling van de toetsingsgrootheid voor het aantal factoren. Als *geschatte* elementen nul zijn, kunnen überhaupt geen asymptotische standaardafwijkingen berekend worden.

d. Een oneigenlijke oplossing kan er op duiden dat het model niet juist of onbruikbaar is. Vaak zal men verwachten dat iedere variabele een vrij grote, onafhankelijke meetfout bevat en dus dat de elementen van $\Psi$ (aanzienlijk) groter dan nul zijn.

Een drietal methoden om oneigenlijke oplossingen aan te pakken zullen wij hier bespreken.

Ten eerste ligt het voor de hand om de variabelen die niet-positieve uniciteiten schijnen te hebben, gewoon weg te laten en met een kleiner aantal variabelen opnieuw te beginnen. Een nadeel hiervan is een verlies aan informatie, waardoor de factoren misschien slechter gedetermineerd (lagere GCW) zullen worden. Verder heeft men niet onderzocht of een weggelaten variabele feitelijk een heel kleine uniciteit heeft en wel degelijk in het model hoort.

LAWLEY & MAXWELL (1971) raden aan om, wanneer men op een oneigenlijke oplossing stuit, een nieuw model te gebruiken. Men neemt dan aan dat bekend is dat de betrokken uniciteiten nul zijn en gaat over tot het schatten en toetsen in dit nieuwe model. De numerieke moeilijkheden zijn hiermee verholpen: als r van de werkelijke $\psi_{ii}$ nul zijn in een m-factormodel (d.w.z. in een ontbinding van een covariantiematrix voor m factoren), dan geldt een (m-r)-factormodel voor de (p-r) × (p-r)-matrix van partiële covarianties voorwaardelijk op de corresponderende r variabelen. Het ligt voor de hand om dit model te schatten met behulp van de steekproef-covariantiematrix die we met $\underline{S}_{11\cdot 2}$ noteren (1,2 duidt op de partioniëring in p-r en r variabelen). Met behulp van de resultaten van 3.3 kan men inzien dat deze methode tevens de maximum likelihoodschattingen voor het *oorspronkelijke* model oplevert: want in dit model hangt de likelihoodfunctie af van de kleinste eigenwaarde van $\Psi^{-\frac{1}{2}}\underline{S}\Psi^{-\frac{1}{2}}$ en deze is continu in $\Psi$. Als $\Psi$ convergeert naar de matrix $\begin{bmatrix} \Psi_1 & 0 \\ \hline 0 & 0 \end{bmatrix}$, naderen zijn eigenwaarden tot de

eigenwaarden van $\Psi_1^{-\frac{1}{2}} \underline{S}_{11 \cdot 2} \Psi_1^{-\frac{1}{2}}$. Tevens heeft dan de likelihood ratio toetsingsgrootheid voor het aantal factoren in het nieuwe model dezelfde waarde als de toetsingsgrootheid in het oorspronkelijke model (d.w.z. waarin $\Psi$ geheel vrij is, mits niet negatief-definiet). Evenwel hebben ze een verschillend aantal vrijheidsgraden (het nieuwe model heeft er meer). Bovendien heeft de toetsingsgrootheid voor het oorspronkelijke model geen asymptotische $\chi^2$-verdeling als één of meer $\psi_{ii}$'s werkelijk nul zijn.

Een heel andere aanpak bieden VAN DRIEL, PRINS & VELDKAMP (1974). Voor hun oplossingsmethode laten zij de eis dat $\Lambda\Lambda'$ en $\Psi$ positief definiet zijn, vallen. Zij schatten met maximum likelihood onder de veronderstelling van normaliteit van de waarnemingen het model $\Sigma = \Xi + \Psi$ met $\Xi$ symmetrisch en van de rang m, en $\Psi$ diagonaal. Blijken de geschatte $\Xi$ en $\Psi$ positief definiet te zijn, dan is er geen verschil met AFA. Is dat voor $\Psi$ evenwel niet zo, dan kijken zij of binnen het betrouwbaarheidsgebied rond $\Psi$ wel niet-negatief definiete $\tilde{\Psi}$ zijn te vinden, waarvan zij dan de dichtstbijzijnde $\tilde{\Psi}$ en de bijbehorende $\tilde{\Xi}$ als schatting kiezen. ($\tilde{\Xi}$ zal dan vaak positief semidefiniet zijn.) Hun methode heeft de volgende voordelen:

a. Numeriek is het schatten voor hun model veel eenvoudiger dan voor AFA omdat zij geen rekening hoeven te houden met grenzen.
b. In statistisch opzicht heeft men het voordeel altijd standaardafwijkingen voor de schatters van $\Psi$ te kunnen schatten en altijd de waarde van m te kunnen toetsen.
c. In principe kan nu onderscheid gemaakt worden tussen oneigenlijke oplossingen veroorzaakt doordat een model geldt met sommige $\psi_{ii}$ (dichtbij) nul en doordat geen model past met de onderhavige m.
d. Ervaring met deze methode wijst erop, dat hij soms veel betere resultaten geeft. Prins en Van Driel contrueerden als voorbeeld een 2-factormodel met $\psi_{11} = 0{,}999$. Hun methode geeft, met 700 waarnemingen, een schatting van 1,003. De dichtsbijzijnde toegelaten waarde is 1, een heel aardige schatting van $\psi_{11}$. De maximum likelihoodschatting voor AFA geeft 0, dus een erg slechte oneigenlijke oplossing.

Voor de praktijk van de factoranalyse zegt hun voorbeeld misschien niet zo veel omdat, in de canonieke oplossing, de tweede kolom van $\Lambda$ praktisch nul is. Zo'n factor zou vrijwel zeker nimmer worden opgemerkt (de

toets op één factor zou in een dergelijke situatie haast nooit tot verwerping leiden) en men zou hebben volstaan met een schatting met slechts één factor. Bij de gegeven data verwerpt de toets op één factor inderdaad niet. De geschatte factor lijkt in de canonieke oplossing sprekend op de eerste factor van hun model.

## 3.7. ASYMPTOTISCHE RAAKHEID VAN DE MAXIMUM LIKELIHOODSCHATTERS VOOR HET ALGEMENE MODEL ZONDER VERONDERSTELLING VAN NORMALITEIT

ANDERSON & RUBIN (1956) bewijzen de asymptotische normaliteit van de -op multinormaliteit gebaseerde- maximum likelihoodschatters voor het algemene model als de veronderstelling van multinormaliteit van de waarnemingen niet is vervuld. Dan moet echter aan een aantal regulariteitsvoorwaarden zijn voldaan, waarvan de belangrijkste identificeerbaarheid van $\Psi$ gegeven m, is (zie 2.11). Het bewijs berust op de asymptotische normaliteit van de steekproefcovariantiematrix. Dit resultaat impliceert de asymptotische raakheid van de schatters.

GILL (1977) geeft een ander bewijs van de asymptotische raakheid, dat, in tegenstelling tot dat van Anderson en Rubin, ook geldig is wanneer sommige uniciteiten nul zijn (de maximum likelihoodschattingen in zo'n model kan men verkrijgen via een methode van Lawley en Maxwell, zoals in 3.6 wordt beschreven, waar ook verdere discussie omtrent dergelijke oneigenlijke modellen staat). Ook dit bewijs vereist identificeerbaarheid van het model. Het berust op het volgende: zij $M(\underline{S})$ de vector van maximum likelihoodschatters als functie van de steekproefcovariantiematrix $\underline{S}$. Gill toont aan dat M continu is en dat in het punt $S = \Sigma$, de echte (populatie) covariantiematrix, $M(\Sigma) = (\Lambda,\Psi)$. Convergentie (in kans of bijna zeker) van $\underline{S}$ naar $\Sigma$ garandeert dan hetzelfde soort convergentie van $M(\underline{S})$ naar de echte parameters.

Asymptotische raakheid zonder assumptie van normaliteit is een, zij het wat vage, geruststelling voor hen die twijfelen aan de gerechtvaardigdheid van een normaliteitsveronderstelling voor hun probleem.

## 3.8. VERSLAG VAN ENIGE SIMULATIE-EXPERIMENTEN*

Multivariaat normaal verdeelde waarnemingen werden gegenereerd volgens een negental factormodellen (alle AFA) en onderworpen aan een aantal analysemethoden die op het Mathematisch Centrum voorhanden waren, te weten de vijf factoranalysemethoden van het programmapakket SPSS (NIE e.a. (1975)), geheten PA1, PA2, RAO, ALPHA en IMAGE (vgl. 2.3), een programma dat de directe oplossingsmethode voor het Jøreskogmodel uitvoert (STATAL (1976)), geheten JORFA en een programma dat de maximum likelihoodmethode volgens Jøreskog voor het algemene model uitvoert (STATAL (1976)), geheten UMLFA. Ten tijde van deze simulatie was een GFA-programma nog niet gereed.

De gebruikte factormodellen (d.w.z. covariantiematrices $\Sigma = \Lambda\Lambda' + \Psi$) waren gegenereerd volgens vijf typen, door ladingen en specifieke varianties apart te genereren volgens bepaalde verdelingen. De ladingen werden per rij op grond van de corresponderende specifieke variantie met een constante vermenigvuldigd zodat een covariantiematrix $\Sigma$ met varianties gelijk aan 1 ontstond. Specifieke varianties werden met kans $\frac{4}{5}$ uit een uniforme verdeling op $[0, \frac{1}{5}]$ en met kans $\frac{1}{5}$ uit een uniforme verdeling op $[\frac{1}{5}, 1]$ getrokken.

De vijf typen van ladinggeneratie waren:

1. hom (-1, 1): alle ladingen uniform verdeeld op [-1, 1]
2. ω : alle ladingen verdeeld volgens fig. 3.8.1.

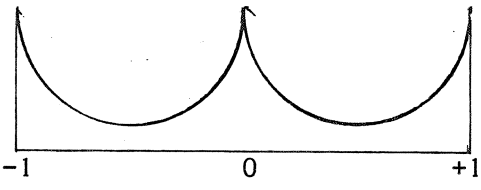



fig. 3.8.1
Drietoppige verdeling op {-1,0,1}

3. nul : alle ladingen gelijk nul, wat resulteert in $\Sigma = I$
4. simpel : $\Lambda = \begin{pmatrix} I_m \\ \cdots \\ I_m \\ \cdots \end{pmatrix}$, oftewel $\lambda_{ij} = \delta_{(i-1) \bmod m, j-1}$

*Voor meer details kunt U bij de schrijvers terecht.

5. cyclisch : $\Lambda = \begin{pmatrix} 1 & 2 \ldots . & m \\ 2 & 3 \ldots . & m+1 \\ \vdots & & \vdots \\ p-1 & p \ldots . & m-2 \\ p & 1 \ldots . & m-1 \end{pmatrix}$ , oftewel $\lambda_{ij} = 1 + (i+j-2) \bmod p$.

Tenslotte werden voor ieder model waarnemingen gegenereerd. Het was de bedoeling na te gaan of bepaalde ladingenstructuren moeilijker te ontdekken zijn dan andere; het bleek, dat de cyclische structuur moeilijkheden geeft bij het determineren van het aantal factoren: de kolommen van $\Lambda$ lijken teveel op elkaar. Tussen de andere structuren waren geen duidelijke verschillen te constateren.

De methoden werden beoordeeld op hun bepalen van het aantal factoren, alsook op hun schatting van $\Lambda$ door $\underline{L}$ (waarbij $\Lambda$ en $\underline{L}$ aan een varimaxrotatie werden onderworpen om ze elementsgewijs vergelijkbaar te maken), van $\Psi$ door $\underline{P}$ en de reproductie van de steekproefcorrelatiematrix $\underline{S}$ door de geschatte $\underline{LL}' + \underline{P}$.

Wanneer verschillende methoden tot een gelijk aantal factoren kwamen, waren de verschillen op de overige punten vrij klein. De precieze resultaten zijn daarom hier niet opgenomen. In het algemeen valt het volgende te concluderen. De methode IMAGE kwam, hoewel hij doorgaans 1½ maal zoveel factoren vond als de andere, tot een slechte reproductie van de correlatiematrix. De iteratieve methoden (PA2, RAO, ALPHA en UMLFA) waren qua schatting en reproductie, *bij gelijk aantal factoren*, beter dan de eenstapsmethoden (PA1 en JORFA). JORFA was meestal beter dan PA1; tussen de iteratieve methoden was weinig verschil. De methode UMLFA die we beneden op andere gronden aanbevelen gaf hier altijd zeer goede resultaten.

De methoden zijn het gemakkelijkst te beoordelen aan de hand van het aantal factoren. PA1, PA2, RAO en ALPHA gebruiken alle het criterium van het aantal eigenwaarden van de steekproefcorrelatiematrix groter dan 1. In 2.10 is al besproken dat dit geen erg gelukkig criterium is. In de methode RAO wordt wel een toets uitgevoerd maar het resultaat wordt genegeerd! JORFA en UMLFA schatten het aantal factoren door een herhaalde statistische toets, hier op 5%-niveau. De toets in JORFA blijkt even onbetrouwbaar te zijn als het eigenwaardecriterium gebruikt in het SPSS-pakket. Men kan ook het aantal factoren in het Jøreskog-model schatten

door de eigenschap $\theta < 1$ te gebruiken; men neemt het laagste aantal waarbij $\underline{t}$ (schatter van $\theta$) $< 1$. Deze methode bleek niet beter. Wel moet worden opgemerkt, dat bij deze experimenten geen van de modellen aan het Jøreskog-model voldeden.

In deze experimenten komt UMLFA duidelijk naar voren met zijn zeer vaak juiste bepaling van het aantal factoren, hetgeen des te bemoedigender is, gezien de kleine verhouding: aantal waarnemingen/aantal variabelen. Bij de eerste drie experimenten komt UMLFA tot een oneigenlijke oplossing, gezien de minimale uniciteit in deze gevallen (resp. 0,0050; 0,0084; 0,0158) is dit niet ten onrechte.

Een zwak punt van de programma's JORFA en RAO is hun onmogelijkheid de toets op nul factoren uit te voeren. Alle methoden, behalve UMLFA, komen tot lachwekkende resultaten als het aantal factoren nul is (experiment 9).

IMAGE komt, zoals gezegd, systematisch tot een veel te hoog aantal factoren, terwijl deze methode ook vaak helemaal niet tot een oplossing kan komen door numerieke gevoeligheid.

De iteratieve methoden PA2, RAO en ALPHA convergeren zeer vaak niet binnen het SPSS standaard aantal van 5 iteraties; zelfs als dit aantal verhoogd wordt tot 25 biedt dat dikwijls geen soelaas. Met dit verhoogde aantal kwamen deze methoden vaak tot oneigenlijke oplossingen, die de vorm aannamen van communaliteitenschattingen groter dan 1. PA2 en RAO stoppen in arren moede terwijl ALPHA willens en wetens doorgaat. Wat men dan met de resultaten moet is een raadsel. UMLFA convergeert zeer snel en bevat een mechanisme dat speciale maatregelen neemt wanneer dergelijke Heywood-gevallen optreden (volgens de methode van Lawley en Maxwell, zie 3.6). Een nette schatting van $\Lambda$ en $\Psi$ wordt gegeven met sommige $P_{ii}$'s gelijk aan nul; in deze gevallen waren de werkelijke $\psi_{ii}$'s ook heel dicht bij nul.

CONCLUSIE. Gebruik zo mogelijk UMLFA, zeker als het aantal factoren onbekend is. Als een iteratieve methode te duur is, lijkt JORFA iets beter dat PA1, maar het bepalen van het aantal factoren vereist, voor beide methoden, veel intuïtie. Gebruik nooit IMAGE.

Tabel 3.8.1 biedt een overzicht van de simulatie-experimenten. Het aangegeven aantal factoren is bij JORFA en UMLFA gevonden door een toets op het aantal factoren. De SPSS-programma's maken gebruik van een eigenwaardecriterium, behalve IMAGE, dat een eigen criterium heeft. Bij deze programma's is het dus mogelijk dat het aangegeven aantal factoren door een toets verworpen wordt.

tabel 3.8.1

Overzicht van simulatie-experimenten

| experimentnummer | aantal factoren | aantal variabelen | aantal waarnemingen | ladingenstructuur | minimale uniciteit | methode | aangegeven aantal factoren | opmerkingen |
|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 10 | 50 | hom(-1,1) | 0,0050 | PA1 | 3 | |
| | | | | | | PA2 | 3 | niet convergent na 5 iteratie-slagen |
| | | | | | | RAO | 3 | niet convergent na 5 iteratie-slagen; toets verwerpt |
| | | | | | | ALPHA | 3 | niet convergent na 5 iteratie-slagen |
| | | | | | | IMAGE | 5 | |
| | | | | | | JORFA | 3 | |
| | | | | | | UMLFA | 3 | oneigenlijke oplossing |
| 2 | 3 | 10 | 50 | $\omega$ | 0,0084 | PA1 | 2 | |
| | | | | | | PA2 | 2 | |
| | | | | | | RAO | 2 | niet convergent na 25 iteratie-slagen |
| | | | | | | ALPHA | 2 | |
| | | | | | | IMAGE | 4 | |
| | | | | | | JORFA | 3 | |
| | | | | | | UMLFA | 3 | oneigenlijke oplossing |

| experimentnummer | aantal factoren | aantal variabelen | aantal waarnemingen | ladingenstructuur | minimale uniciteit | methode | aangegeven aantal factoren | opmerkingen |
|---|---|---|---|---|---|---|---|---|
| 3 | 3 | 10 | 50 | cydisch | 0,0198 | PA1 | 2 | |
| | | | | | | PA2 | 2 | |
| | | | | | | RAO | 2 | communaliteiten > 1 na 8 iteratie-slagen |
| | | | | | | ALPHA | 2 | niet convergent na 25 iteratie-slagen |
| | | | | | | IMAGE | 5 | |
| | | | | | | JORFA | 2 | |
| | | | | | | UMLFA | 2 | oneigenlijke oplossing |
| 4 | 3 | 10 | 50 | simpel | 0,0039 | PA1 | 3 | |
| | | | | | | PA2 | 3 | niet convergent na 5 iteratie-slagen |
| | | | | | | RAO | 3 | niet convergent na 5 iteratie-slagen |
| | | | | | | ALPHA | 3 | niet convergent na 5 iteratie-slagen; communaliteiten > 1 |
| | | | | | | IMAGE | 5 | |
| | | | | | | JORFA | 3 | |
| | | | | | | UMLFA | 3 | |
| 5 | 5 | 15 | 200 | simpel | 0,0009 | PA1 | 5 | |
| | | | | | | PA2 | 5 | communaliteiten > 1 na 9 itera-tieslagen |
| | | | | | | RAO | 5 | niet convergent na 25 iteratie-slagen; toets verwerpt |
| | | | | | | ALPHA | 5 | |
| | | | | | | IMAGE | 7 | |
| | | | | | | JORFA | - | toets verwerpt 5,6,7 factoren eigenwaardecriterium geeft 5 factoren |
| | | | | | | UMLFA | 5 | |

| experimentnummer | aantal factoren | aantal variabelen | aantal waarnemingen | ladingenstructuur | minimale uniciteit | methode | aangegeven aantal factoren | opmerkingen |
|---|---|---|---|---|---|---|---|---|
| 6 | 5 | 15 | 50 | simpel | 0,0009 | PA1 | 5 | |
| | | | | | | PA2 | 5 | communaliteiten > 1 na 3 iteratieslagen |
| | | | | | | RAO | 5 | niet convergent na 25 iteratieslagen; toets verwerpt |
| | | | | | | ALPHA | 5 | 2 communaliteiten > 1 |
| | | | | | | IMAGE | - | geen oplossing daar de "correlatiematrix bijna singulier" is |
| | | | | | | JORFA | - | toets verwerpt 5 en 6 factoren. eigenwaardecriterium levert 5 factoren. |
| | | | | | | UMLFA | 5 | |
| 7 | 4 | 75 | 150 | ω | 0,0015 | PA1 | 4 | |
| | | | | | | PA2 | 4 | |
| | | | | | | RAO | 4 | niet convergent na 5 iteratieslagen; toets verwerpt zeer duidelijk |
| | | | | | | ALPHA | 4 | |
| | | | | | | IMAGE | - | geen oplossing daar de "correlatiematrix bijna singulier" is |
| | | | | | | JORFA | - | toets en eigenwaardecriterium verwerpen 4 en 5 factoren |
| | | | | | | UMLFA | 4 | |
| 8 | 1 | 15 | 50 | hom(-1,1) | 0,0009 | PA1 | 2 | |
| | | | | | | PA2 | 2 | communaliteiten > 1 na 2 iteratieslagen |
| | | | | | | RAO | 2 | niet convergent na 25 iteratieslagen; toets verwerpt |
| | | | | | | ALPHA | 2 | niet convergent met communaliteiten > 1 |
| | | | | | | IMAGE | - | geen oplossing daar de "correlatiematrix bijna singulier" is |
| | | | | | | JORFA | 1 | eigenwaardecriterium levert minstens 4 factoren |
| | | | | | | UMLFA | 1 | |

| experimentnummer | aantal factoren | aantal variabelen | aantal waarnemingen | ladingenstructuur | minimale uniciteit | methode | aangegeven aantal factoren | opmerkingen |
|---|---|---|---|---|---|---|---|---|
| 9 | 0 | 15 | 50 | nul | 1,0000 | PA1 | 7 | |
| | | | | | | PA2 | 7 | niet convergent na 25 iteratie-slagen |
| | | | | | | RAO | 7 | niet convergent na 25 iteratie-slagen, toets op 1 factor verwerpt niet, toets op 0 factoren niet mogelijk |
| | | | | | | ALPHA | 7 | niet convergent na 25 iteratie-slagen |
| | | | | | | IMAGE | 8 | |
| | | | | | | JORFA | 1 | toets op 0 factoren is niet mogelijk |
| | | | | | | UMLFA | 0 | |

REFERENTIES:

[1] ANDERSON, T.W., *Asymptotic theory for principal component analysis*, Ann. of Math. Stat. 34, (1963) p. 122.

[2] ANDERSON, T.W. & H. RUBIN, *Inference in factor analysis*, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 5, (1956) p. 111.

[3] ANDREWS, D.F., R. GNANADESIKAN & J.L. WARNER, *Methods for assesing multivariate normality*, Multivariate analysis III, Academic Press, New York, (1973) p. 95.

[4] ANSCOMBE, F.J. & J.W. TUKEY, *The examination of residuals*, Technometrics 5, (1963), p. 141.

[5] BETHLEHEM, J. (editor), *Statal reference manual*, Mathematical Centre, Amsterdam, (1976).

[6] DIXON, W.J. (editor), *BMD: Biomedical computer programs*, Univ. of California Press, Los Angeles, (1975).

[7] DRIEL, P.O., van, *Description of the malifa (maximum likelihood factor analysis) program*, Nat. Lab. Computer Note nr. 1975/1, Philips Research Laboratories, Eindhoven, (1975)

[8] DRIEL, O.P., van, H.J. PRINS & G.W. VELDKAMP, *Estimating the parameters of the factor analysis model without the usual constraints of positive definiteness*, Proceedings of the Symposium on Computational Statistics, Vienna, (1974), p. 255.

[9] EMMET, W.G., *Factor analysis by Lawley's method of maximum likelihood*, Brit. J. Psych. 2, (1949), p. 90.

[10] FUKUTOMI, K., *On the adequacy of factor extractions*, TRU Mathematics 9, (1973), p. 119.

[11] FULLER, E.J. & W.J. HEMMERLE, *Robustness of the maximum likelihood estimation procedure in factor analysis*, Psychometrika 31, (1966), p. 255.

[12] GILL, R.D., *Constistency of maximum likelihood estimators of the*

*factor analysis model, when observations are not multivariate normally distributed,* In: Recent developments in statistics (editor: J.R. BARRA), North-Holland Publishing Company, Amsterdam, (1977).

[13] GUTTMAN, L., *Image theory for the structure of quantitive variates,* Psychometrika 18, (1953), p. 277.

[14] GUTTMAN, L., *Some necessary conditions for common - factor analysis,* Psychometrika 19, (1954), p. 149.

[15] GUTTMAN, L., *The determinacy of factor score matrices with implications for five other basic problems of common - factor theory,* Brit. J. Stat. Psych. 8, (1955), p. 65.

[16] HARMAN, H.H., *Modern factor analysis,* Univ. of Chicago Press, Chicago, (1960).

[17] HEGAZY, Y.A.S. & J.R. GREEN, *Some new goodness-of-fit tests using order statistics,* J. Appl. Stat. 24, (1975), p. 299.

[18] HEMELRIJK, J. & J. KRIENS, *Leergang Besliskunde, deel 3,* M.C. Syllabus 1.3, Mathematisch Centrum, Amsterdam, (1967).

[19] HENDRICKSON, A. & P. WHITE, *A quick method for rotation to oblique simple structure,* Brit. J. Stat. Psych. 19, (1969), p. 65.

[20] JENNRICH, R.I., *Simplified formulae for standard errors in maximum likelihood factor analysis,* Brit. J. Math. Stat. Psych. 27, (1974), p. 122.

[21] JORESKOG, K.G., *On the statistical treatment of residuals in factor analysis,* Psychometrika 27, (1962), p. 51.

[22] JORESKOG, K.G., *Statistical estimations in factor analysis,* Almqvist & Wiksell, Uppsala, (1963).

[23] JORESKOG, K.G., *Some contributions to maximum likelihood factor analysis,* Psychometrika 32, (1967), p. 443.

[24] JORESKOG, K.G., *Efficient estimation in image facor analysis,* Psychometrika 34, (1969), p. 51.

[25] JORESKOG, K.G., *A general method for analysis of covariance structures*, Biometrika 57, (1970), p. 239.

[26] KAISER, H.F., *The varimax criterion for analytic rotation in factor analysis*, Psychometrika 23, (1958), p. 187.

[27] KAISER, H.F., *Image analysis*, Problems in measuring chance, Univ. of Wisconsin Press, (1963).

[28] KAISER, H.F. & J. CAFFREY, *Alpha factor analysis*, Psychometrika 30, (1965), p. 1.

[29] LAWLEY, D.N. & A.E. MAXWELL, *Factor analysis as a statistical method*, American Elsevier Publishing Company, New York, (1971).

[30] LORD, F.M. & M.R. NOVICK, *Statistical theories of mental test scores*, Addison-Wesley, Reading, (1968).

[31] MAXWELL, A.E., *Recent trends in factor analysis*, J. Roy. Stat. Soc. Series A 124, (1961), p. 49.

[32] MAXWELL, A.E., *Calculating maximum likelihood factor loadings*, J. Roy. Stat. Soc. Series A 127, (1964), p. 238.

[33] MORRISON, D.F., *Multivariate statistical methods*, McGraw-Hill, New York, (1976).

[34] McDONALD, R.P., Th., *The measurement of factor inderterminacy*, Psychometrika 39, (1974), p. 203.

[35] MULAIK, S.A., *The foundations of factor analysis*, McGraw-Hill, New York, (1972).

[36] MULAIK, S.A., *Comments on "The measurement of factor indeterminacy"*, Psychometrika 41, (1976), p. 249.

[37] NIE, N.H. et al., *SPSS: Statistical package for the social sciences*, McGraw-Hill, New York, (1975).

[38] RAO, C.R., *The use and interpretation of principal component analysis in applied research*, Sankya 26, (1964), p. 329.

[40] RAO, C.R., *Estimation and tests of significance in factor analysis*, Psychometrika 20, (1955), p. 93.

[41] REIERSØL, O., *On the identifiability of parameters in Thurstone's multiple factor analysis*, Psychometrika 15, (1950), p. 121.

[42] SCHOENEMAN, P.H. & M. WANG, *Some new results on factor inderterminacy*, Psychometrika 37, (1972), p. 61.

[43] SCOTT ARMSTRONG, J., *Derivation of theory of factor analysis or Tom Swift and his electric factor analysis machine*, The American Statistician 21, (1967), p. 17.

[44] SILVEY, S.D., *Statistical inference*, Penguin Books, Haimondsworth, (1970).

[45] SLATER, P. & E. BENNET, *The development of spatial judgement and its relations to some educational problems*, J. Occ. Psych. 17, (1943), p. 139.

[46] SWAIN, A.J., *A class of factor analysis estimation procedures with common asymptotic sampling properties*, Psychometrika 40, (1975), p. 315.

[47] TATSUOKA, M.M., *Multivariate analysis*, Techniques for educational and psychological research, Wiley, New York, (1971).

[48] THURSTONE, L.L., *Multiple factor analysis*, Univ. of Chicago Press, Chicago, (1947).

[49] UBERLA, K., *Factorenanalyse*, Springer, Heidelberg, (1968).

[50] WILKINSON, J.H., *The algebraic eigenvalue problem*, Clarendon Press, Oxford, (1965).

[51] WILSON, E.B. & J. WORCESTER, *The resolution of six texts into three general factors*, Proc. Nat. Acad. Sci., Washington 25, (1939), p. 73.

REGISTER